

Model-based pretreatment for correlated datasets,
Application to linear regression and missing values.
Real datasets from steel industry

Clément THERY

July 25, 2014

To my sons

Résumé

Les travaux effectués durant cette thèse ont pour but de pouvoir pallier le problème des corrélations au sein des bases de données, particulièrement fréquentes dans le cadre des données industrielles. Une modélisation explicite des corrélations par un système de sous-régressions entre covariables permet de pointer les sources des corrélations et d'isoler certaines variables redondantes.

Il en découle une pré-sélection de variables nettement moins corrélées sans perte significative d'information et avec un fort potentiel explicatif (la préselection elle-même est expliquée par la structure de sous-régression qui est simple à comprendre car uniquement constituée de modèles linéaires).

Un algorithme de recherche de structure de sous-régressions est proposé, basé sur un modèle génératif complet sur les données et utilisant une chaîne MCMC (Monte-Carlo Markov Chain). Ce prétraitement est utilisé pour la régression linéaire à des fins illustratives mais ne dépend pas de la variable réponse et peut donc être utilisé de manière générale pour toute problématique de corrélations.

Par suite, le modèle génératif complet peut être utilisé pour gérer d'éventuelles valeurs manquantes dans les données, tant pour la recherche de structure que pour de l'imputation multiple préalable à l'utilisation de méthodes classiques incompatibles avec la présence de valeurs manquantes. Cela permet également d'estimer les valeurs manquantes et de fournir un intervalle de confiance sur leur estimation. Encore une fois, la régression linéaire vient illustrer l'apport de la méthode qui reste cependant générique et applicable à d'autres contextes tels que le clustering.

Enfin, un estimateur plug-in pour la régression linéaire est proposé pour ré-injecter les variables redondantes de manière séquentielle et donc utiliser toute l'information sans souffrir des corrélations entre covariables.

Tout au long de ces travaux, l'accent est mis principalement sur l'interprétabilité des résultats en raison du caractère industriel du financement de cette thèse.

Le package R intitulé CorReg, disponible sur le CRAN¹ sous licence CeCILL², implémente les méthodes développées durant cette thèse.

Mots clés: Prétraitement, Régression, Corrélations, Valeurs manquantes, MCMC, modèle génératif, Critère Bayésien, sélection de variable, méthode séquentielle.

¹<http://cran.r-project.org>

²<http://www.cecill.info>

Abstract

This thesis was motivated by correlation issues in real datasets, in particular industrial datasets. The main idea stands in explicit modeling of the correlations between covariates by a structure of sub-regression, that simply is a system of linear regression between the covariates. It points out redundant covariates that can be deleted in a pre-selection step to improve matrix conditonning without significant loss of information and with strong explicative potential because this pre-selection is explained by the structure of sub-regression, itself easy to interpret.

An algorithm to find the sub-regression structure inherent to the dataset is provided, based on full generative model and using Monte-Carlo Markov Chain (MCMC) method. This pretreatment is then illustrated on linear regression to show its efficiency but does not depend on a response variable and thus can be used in a more general way with any correlated datasets.

The generative model defined here allows to manage missing values both during the MCMC and then for imputation (for example multiple imputation) to be able to use classical methods that are not compatible with missing datasets. Missing values can be imputed with a confidence interval to show estimation accuracy. Once again, linear regression is used to illustrate the benefits of this method but it remains a pretreatment that can be used in other contexts, like clustering and so on.

Finally a plug-in estimator is defined to get back the redundant covariates sequentially. Then all the covariates are used but the sequential approach act as a protection against correlations.

The industrial motivation of this work define interpretation as a stronghold at each step.

The R package CorReg, is on CRAN³ now under CeCILL⁴ license. It implements the methods created during this thesis.

Keywords: Pretreatment, Regression, Correlations, Missing values, MCMC, generative model, Bayesian Criterion, variable selection, plug-in method,...

³<http://cran.r-project.org>

⁴<http://www.cecill.info>

Résumé substantiel en français

en gros, un article en français

Acknowledgments

Contents

1	The industrial context	9
1.0.1	Steelmaking process	9
1.0.2	Impact of the industrial context	9
2	State of the art	11
2.1	Linear regression	11
2.1.1	Historic interest	11
2.1.2	Simplicity	11
2.1.3	Industrial context	11
2.1.4	Flexibility and future of regression	12
2.2	Ordinary least squares and associated problems	12
2.3	Penalized models	13
2.3.1	Ridge regression	14
2.3.2	LASSO: Least Absolute Shrinkage and Selection Operator	15
2.3.3	Least Angle Regression	15
2.3.4	Elasticnet	16
2.3.5	OSCAR: Octogonal Shrinkage and Clustering Algorithm for Regression	16
2.4	Modeling the parameters	16
2.4.1	CLERE: CLusterwise Effect REgression	16
2.4.2	Spike and Slab	17
2.5	Miscellaneous	17
2.5.1	Principal Component Regression	17
2.5.2	Partial Least Squares (PLS) Regression	17
2.5.3	Non-parametric regression	17
2.5.4	Sliced Inverse Regression	17
2.5.5	General Linear Model (GLM)	17
2.5.6	Classification and Regression Trees (CART)	18
2.5.7	Neural networks	18
2.5.8	Bayesian networks	18
2.6	Choice of model	18
2.6.1	Cross validation	18
2.6.2	Information Criterion	18
2.6.3	stepwise	18
2.6.4	bootstrap	18
2.7	SEM	18
2.8	MCMC	18
2.9	Gibbs	18
2.10	Industrial tools	19
2.11	Multiple Equations	19
2.11.1	SEM and Path Analysis	19

2.11.2	SUR: Seemingly Unrelated Regression	19
2.11.3	SPRING: Structured selection of Primordial Relationships IN the General linear model	19
2.11.4	Selvarchlust: Linear regression within covariates for clustering	19
I	Pretreatment for correlations	20
3	Decorrelating covariates by a generative model	21
3.1	Our proposal: modelisation of the correlations	21
3.2	Graph theory	22
3.3	A by-product model: marginal regression with decorrelated covariates	23
3.4	Strategy of use: pre-treatment before classical estimation/selection methods	24
4	Numerical results with a known structure	25
4.1	Illustration of the tradeoff conveyed by the pre-treatment	25
4.2	Predictive efficiency	26
5	Estimation of the Structure of subregression by MCMC	27
5.1	Bayesian approach	27
5.2	Sub-regression structure in details	27
5.2.1	Modeling the uncorrelated covariates: a full generative approach on $P(\mathbf{X})$	27
5.2.2	Identifiability of the structure	28
5.3	Sub-regression model selection	29
5.3.1	Bayesian criterion for quality	29
5.3.2	Penalization of the integrated likelihood by $P(S)$	29
5.3.3	Some indicators for proximity	30
5.4	Neighbourhood	30
5.4.1	Strategy	30
5.4.2	Active relaxation of the constraints	31
5.5	The walk	31
5.6	Initialization	32
5.6.1	Correlation-based initialization	32
5.6.2	Multiple intialisation	32
5.7	Pruning	33
5.8	The Graphical LASSO	34
5.9	CorReg	34
6	Numerical results on simulated datasets	36
6.1	The datasets	36
6.2	Results on \hat{S}	36
6.3	Results on prediction	38
6.3.1	\mathbf{Y} depends only on covariates in \mathbf{X}^{I_f} (best case for us)	38
6.3.2	\mathbf{Y} depends on all variables in \mathbf{X}	40
6.3.3	\mathbf{Y} depends only on covariates in \mathbf{X}^{I_r} (worst case for us)	42
7	Numerical results on real datasets	44
7.1	Quality case study	44
7.2	Production case study	46

II	Further usage of the structure	48
8	Missing values	49
8.1	Introduction	49
8.2	Estimation of the sub-regressions with missing values	49
8.2.1	The integrated likelihood	49
8.2.2	Likelihood computation optimized	52
8.2.3	Weighted penalty	53
8.3	SEM	53
8.3.1	Our implementation of SEM	53
8.3.2	Stochastic imputation by Gibbs sampling	54
8.3.3	Alternative E step	55
8.4	Missing values in the main regression	55
8.5	Numerical results on simulated datasets	56
8.5.1	Estimation of the sub-regression coefficients	56
8.5.2	Multiple imputation for the main regression	56
8.6	Missing values in real life	56
9	Taking back the residuals	57
9.1	The model	57
9.2	Interpretation and latent variables	57
9.3	Consistency	58
9.4	Numerical results	59
10	CorReg: the concept	60
11	Conclusion and perspectives	61
11.1	Conclusion	61
11.2	Perspective	61
11.2.1	Non-linear regression	61
11.2.2	Pretreatment not only for regression	61
11.2.3	Improved programming	61
11.2.4	Missing values in classical methods	61
11.2.5	Interpretation improvements	62
	References	63

Chapter 1

The industrial context

Abstract: Ce chapitre explique les contraintes industrielles qui ont orienté les travaux pour répondre aux demandes d'ArcelorMittal qui est le commanditaire de ces travaux de recherche.

1.0.1 Steelmaking process

This work takes place in a steel industry context. Steelmaking starts from raw materials to give highly specific products.

parler du process de manière linéaire



faire remarquer la longueur du process et le nombre de paramètres. Faire remarquer les corrélations Conclure avec le côté innovant de la sidérurgie puis transition vers la recherche et donc vers la thèse

1.0.2 Impact of the industrial 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 may 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 plant. 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

Abstract: Rapide aperçu de ce qui existe déjà pour tenter de répondre à notre problématique.

2.1 Linear regression

2.1.1 Historic interest

méthode ancienne et reconnue, remonte aux origines des statistiques, méthode pionnière en prédiction.

2.1.2 Simplicity

Facile à mettre en oeuvre théoriquement, rapide en pratique et présent partout (même dans Excel) Très simple à interpréter, principe intuitif. donne tout de suite l'impact des variables (positif ou négatif) sur la réponse et leur poids (si scaled dataset)

C'est ici qu'on peut mettre le principe de la régression linéaire (image d'un nuage de point et d'une droite qui le traverse)

2.1.3 Industrial context

Manque d'arrière plan statistique besoin de comprendre pour corriger défaut de confiance dans les statistiques d'où besoin accru en interprétation La régression s'y prête bien car déjà connue et utilisée par chacun dans Excel (parfois à tort à travers).

Industrial context is often poor in statistical background and the stakes are frequently very high in terms of financial impact. These two points give strong constraints because methods used has to be accessible for non-statistician in a minimum amount of time and results obtained have to be clearly interpreted (no black-box) because if industrial experts don't understand the result, they will not trust it and then they will not use it. So a powerful tool without interpretation becomes kind of useless in such a context.

Every engineer, even non-statistician use frequently linear regression to seek relationship between some covariates. It is easy to understand, fast to do, it can be done directly in Microsoft Excel that remains the most used software in industry and the software used by engineers to open most of the datasets.

Regression appears to be the basis of industrial statistics so we have chosen to work in this way. As of 2014 Google Scholar proposes more than 3.8 millions of papers related to regression and many of them were cited several thousands times. It is an old strategy well known and with many derivative. It's simplicity facilitates a wide spread usage in industry and other fields of application.

2.1.4 Flexibility and future of regression

Richesse des types de régression et des méthodes d'estimation On peut faire des choses plus pointues en adaptant un peu le modèle mais en conservant la simplicité : Multilevel Regression [Moerbeek et al., 2003, Maas and Hox, 2004, Hox, 1998]

Notations: In the following we note classical (respectively L_2, L_1, L_∞) norms: $\|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^p (\beta_i)^2$, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^p |\beta_i|$ and $\|\boldsymbol{\beta}\|_\infty = \max(|\beta_1|, \dots, |\beta_p|)$. Vectors are in bold characters.

2.2 Ordinary least squares and associated problems

We note the linear regression model:

$$\mathbf{Y}_{|\mathbf{X}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.1)$$

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

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

with variance matrix

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

and without any bias [Saporta, 2006, Dodge and Rousson, 2004]. In fact it is the Best Linear Unbiased Estimator (BLUE). The theoretical MSE is given by

$$E[\text{MSE}(\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{X})] = 0 + \sigma_Y^2 \text{Tr}((\mathbf{X}'\mathbf{X})^{-1}) \quad (2.4)$$

Equation 2.1 has no intercept but can be generalized by adding to \mathbf{X} a first column full of 1. So we don't consider the intercept to simplify notations. In practice, an intercept is added by default.

Ordinary Least Squares find a p -dimensional hyperplane that minimizes the distance with each individual (\mathbf{X}_i, Y_i) . It can be written

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \right\} \quad (2.5)$$

So estimation of \mathbf{Y} by OLS can be viewed as a projection onto the linear space spanned by the regressors \mathbf{X} as shown in figure 2.1.

Estimation of $\boldsymbol{\beta}$ requires the inversion of $\mathbf{X}'\mathbf{X}$ which will be ill-conditioned or even singular if some covariates depend linearly from each other. For a fixed number n of individuals, conditioning of $\mathbf{X}'\mathbf{X}$ 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)
- The correlations within the covariates: strongly correlated covariates give bad-conditioning and increase variance of the estimators .

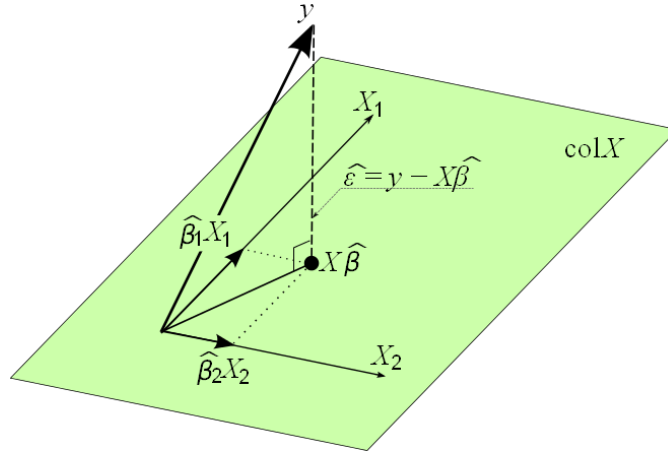


Figure 2.1: Multiple linear regression with Ordinary Least Squares. Public domain image.

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 to have more individuals than covariates ($n \geq p$). When matrices are not invertible, classical packages like the function `lm` of R base package [R Core Team, 2014] use the Moore-Penrose pseudoinverse [Penrose, 1955] to generalize OLS.

Last but not least, Ordinary Least Squares is unbiased but if some β_i are null (irrelevant covariates) the corresponding $\hat{\beta}_i$ will only asymptotically tend to 0 so the number of covariates in the estimated model remains p . This is a major issue because we are searching for a statistical tool able to work without a priori on a big dataset containing many irrelevant datasets. Pointing out some relevant covariate and how they impact the response really is the main goal here. We will need a variable selection method one moment or another. It could be as a pretreatment, during coefficient estimation or by post-treatment.

Running example: we look at a simple case with $p = 5$ variables defined by four independent scaled Gaussian $\mathcal{N}(0, 1)$ named $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5$ 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 two *scenarii* for \mathbf{Y} with $\boldsymbol{\beta} = (1, 1, 1, 1, 1)$ and $\sigma_Y \in \{10, 20\}$. So there is no intercept (can be seen as a null intercept). It is clear that $\mathbf{X}'\mathbf{X}$ will become more ill-conditioned as σ_3 gets smaller.

Figure 2.2 shows the theoretical MSE obtained on $\hat{\boldsymbol{\beta}}$ with OLS. These results are based on equation 2.3, we show the mean obtained after 100 experiences computed on our running example.

The R^2 stands for:

$$R^2 = 1 - \frac{\sum_i (y_i - x_i \beta)^2}{\sum_i (y_i - \bar{y})^2} \quad (2.6)$$

where $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$

Many other estimation methods were created to obtain better estimations by playing on the bias/variance tradeoff or by making additional hypothesis. To have an easier comparison, we also look at the empiric MSE obtained on $\hat{\boldsymbol{\beta}}$.

Image sur MSE réels

2.3 Penalized models

We have seen that OLS is the Best linear Unbiased Estimator for $\hat{\boldsymbol{\beta}}$, meaning that it has the minimum variance. But it remains possible to play with the bias/variance tradeoff to reduce



Figure 2.2: Evolution of theoretical Mean Squared error on $\hat{\beta}$ with the strength of the correlations for various sample sizes and strength of regression. $p = 5$ covariates.

the variance by adding some bias. The underlying idea is that a small bias and a small variance could be preferred to a huge variance without bias. Many methods do this by a penalization on $\hat{\beta}$. Some of them propose an effective variable selection.

2.3.1 Ridge regression

Ridge regression [Hoerl and Kennard, 1970, Marquardt and Snee, 1975] proposes a biased estimator for β that can be written in terms of a parametric L_2 penalty:

$$\hat{\beta} = \operatorname{argmin} \left\{ \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_2^2 \leq \lambda \text{ with } \lambda > 0 \quad (2.7)$$

But this penalty is not guided by the correlations. It introduce an additional parameter λ to choose for the whole dataset whereas correlations may concern only some of the covariates with several intensities.

The solution of the ridge regression is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} - \lambda \mathbf{I}_n)^{-1} \mathbf{X}'\mathbf{Y} \quad (2.8)$$

and we see in this equation that a global modification of $\mathbf{X}'\mathbf{X}$ is done for a given λ . Methods does exist to automatically choose a good value for λ [Cule and De Iorio, 2013, Er et al., 2013] and a R package called RIDGE is on CRAN [Cule, 2014]. We have computed the same experiment as in previous figure but with the ridge package instead of OLS. It is clear that the ridge regression is efficient in variance reduction (it is what it is built for).

Moreover, like OLS, coefficients tend to 0 but don't reach 0 so it gives difficult interpretations for large values of p . Ridge regression is efficient to improve conditionning of the estimator but gives no clue to the origin of ill-conditioning and keep irrelevant covariates. It remains a good candidate for prediction-only studies.

2.3.2 LASSO: Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO [Tibshirani, 1996, Tibshirani et al.,]) consists in a shrinkage of the regression coefficients based on a λ parametric L_1 penalty to obtain zeros in $\hat{\beta}$ instead of the L_2 penalty of the ridge regression:

$$\hat{\beta} = \operatorname{argmin} \left\{ \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_1 \leq \lambda \text{ with } \lambda > 0 \quad (2.9)$$

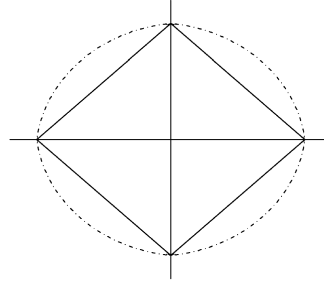


Figure 2.3: Geometric view of the Penalty for the LASSO

détailler le graphique

Here again we have to choose a value for λ . 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. In practice, we know that the LASSO faces consistency issues when confronted to correlated covariates [Zhao and Yu, 2006]. When two covariates are correlated, it tends to keep only one of them. For example, if two covariates are equal and have the same effect, the LASSO will keep only one of them. As explained earlier, variable selection is a real stake for us but is necessary to have a good interpretation. The LASSO does not distinguish a covariate not selected because it is totally redundant with another that was selected from an irrelevant covariate. And that is a problem. This consistency issue is illustrated in section 9.3.

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 the consistency issues remains because it is still the same model. Only the choices of λ differ.

It is notable that the main goal of the LASSO is to select some covariate, thus the penalization is just a mean to achieve selection. But estimation of $\hat{\beta}$ can be improved by a second estimation with OLS based only on selected covariates [Zhang and Shen, 2010].

2.3.3 Least Angle Regression

Schéma de la géométrie qui motive le nom de la méthode

sortie graphique du package lars avec les chemins d'évolution des coefficients.

Explication du choix par validation croisée et de la surcouche OLS

Résultats comparatifs en modifiant le A du running exemple pour voir que validation croisée est mieux que le BIC dans ce cas là (au passage, prendre un cas non consistant).

2.3.4 Elasticnet

Elastic net [Zou and Hastie, 2005] is a method developed to be a compromise between Ridge regression and the LASSO by mixing both L_1 and L_2 penalties:

$$\begin{aligned} \hat{\beta} = (1 + \lambda_2) \operatorname{argmin} \{ & \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \} \text{ subject to} \\ & (1 - \alpha) \| \beta \|_1 + \alpha \| \beta \|_2^2 \leq t \text{ for some } t \end{aligned} \quad (2.10)$$

where $\alpha = \frac{\lambda_2}{(\lambda_1 + \lambda_2)}$.

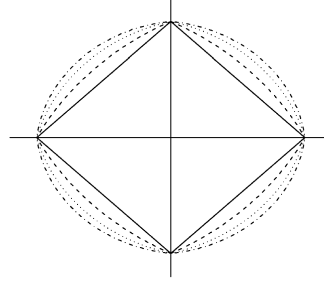


Figure 2.4: Geometric view of the Penalty for elasticnet

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.

2.3.5 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).

$$\hat{\beta} = \operatorname{argmin}_{\beta} \| \mathbf{Y} - \mathbf{X}\beta \|_2^2 \text{ subject to } \sum_{j=1}^p |\beta_j| + c \sum_{j < k} \max(|\beta_j|, |\beta_k|) \leq \lambda \quad (2.11)$$

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 .

détailler le graphique

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.

2.4 Modeling the parameters

2.4.1 CLERE: CLusterwise Effect REgression

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

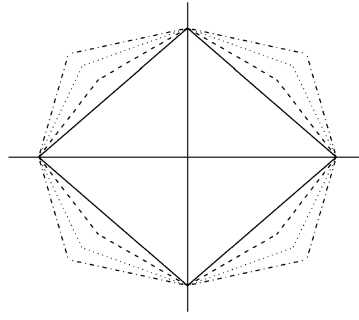


Figure 2.5: Geometric view of the Penalty for OSCAR

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 $\hat{\beta}$.

2.4.2 Spike and Slab

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.

Mettre un dessin du spike et du slab

2.5 Miscellaneous

2.5.1 Principal Component Regression

[Jackson, 2005]

2.5.2 Partial Least Squares (PLS) Regression

[Abdi, 2003, Geladi and Kowalski, 1986]

2.5.3 Non-parametric regression

[Eubank, 1999, Hardle, 1990] Non conforme aux exigences d'interprétabilité

2.5.4 Sliced Inverse Regression

It is a semi-parametric approach that could be seen as easier to interpret than general non-parametric regression. But it is not sufficient in terms of ease of use (after estimation) for non-statisticians. [Li, 1991, Saracco et al., 1999]

2.5.5 General Linear Model (GLM)

[Kiebel and Holmes, 2003, Wickens, 2004, Nelder and Baker, 1972, McCullagh and Nelder, 1989] Generalise le modèle linéaire, assez classique dans les outils statistiques, mais pas assez immédiat en interprétation par rapport au modèle linéaire classique ou bien aux arbres de décision (transition vers la section suivante sur CART)

2.5.6 Classification and Regression Trees (CART)

[Breiman, 1984] Bonne méthode pour l'industrie Son problème principale est le cas linéaire... donc bon complément à notre modèle. CorReg propose un outil de mise en oeuvre rapide et avec légendes ainsi qu'une fonctionnalité de sélection pour mieux analyser les problématiques de corrélations. Gros avantage: les données mixtes

2.5.7 Neural networks

Neural networks [Fausett, 1994] seems to be really powerful but it has a predictive only goal and the model obtained can't be interpreted easily (and can't be interpreted at all if too complex). So it does not correspond to our needs. Interpretation remains our first goal and prediction comes far behind.

2.5.8 Bayesian networks

[Heckerman et al., 1995, Jensen and Nielsen, 2007, Friedman et al., 2000] Bayesian networks are quite good in terms of interpretation but suffer from great dimension and require to transform the dataset arbitrary (discretisation), that imply a loss of information and usage of a priori (that is explicitey not suitable in our industrial context).

2.6 Choice of model

2.6.1 Cross validation

[Kohavi et al., 1995, Arlot et al., 2010]

2.6.2 Information Criterion

[Lebarbier and Mary-Huard, 2006]

2.6.3 stepwise

[Seber and Lee, 2012, Miller, 2002]

2.6.4 bootstrap

[Efron, 1979, Efron and Tibshirani, 1994]

2.7 SEM

2.8 MCMC

[Gilks et al., 1996, Chib and Greenberg, 1995, Roberts and Rosenthal, 2001]

2.9 Gibbs

[Casella and George, 1992]

2.10 Industrial tools

Linear regression, decision trees, bayesian networks, neural network but without confidence
Presence of softwares sent as non-statistical methods, based on rules (derivatives of decision trees).
metre image du flou vert et rouge avec encadré

2.11 Multiple Equations

2.11.1 SEM and Path Analysis

[Davidson and MacKinnon, 1993, Pearl, 2000, Pearl, 1998, Brito and Pearl, 2006]

2.11.2 SUR: Seemingly Unrelated Regression

[Zellner, 1962]

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

[Chiquet J. and S., 2013]

2.11.4 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]

In this work we propose to adapt this model for linear regression and to use it as a pretreatment on correlated covariates. We provide a 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

Abstract: Nous modélisons explicitement les corrélations entre covariables par un système de régressions linéaires entre covariables. Cela permet une meilleure compréhension des données mais aussi une préselection de variables mettant de côté les variables redondantes pour réduire fortement les corrélations tout en ne perdant que peu d'information. La préselection prend un sens particulier grâce à la structure de sous-régression qui permet de distinguer par suite les variables indépendantes de la variable réponse de celles qui sont juste redondantes mais potentiellement liées à la variable réponse.

3.1 Our proposal: modelisation of the correlations

Let \mathbf{X} be a $n \times p$ matrix of observed covariates and \mathbf{Y} be the $n \times 1$ matrix of the observed response variable. In the following, we note \mathbf{X}^j the j^{th} column of \mathbf{X} and \mathbf{X}^J where $J = \{j_1, \dots, j_k\}$ the $n \times k$ sub-matrix of \mathbf{X} composed by the columns of \mathbf{X} whose indices are in the set J .

We make the hypothesis that \mathbf{X} can be described by a partition $\mathbf{X} = (\mathbf{X}^{I_f}, \mathbf{X}^{I_r})$ given by an explicit structure S where variables in the $n \times p_r$ sub-matrix \mathbf{X}^{I_r} are redundant endogenous covariates resulting from linear sub-regressions based on \mathbf{X}^{I_f} , the $n \times (p - p_r)$ sub-matrix of free (mutually independent) exogenous covariates. So we model the correlations by $P(\mathbf{X}^{I_r} | \mathbf{X}^{I_f})$ with \mathbf{X}^{I_f} orthogonal covariates.

The structure S of p_r sub-regressions within correlated covariates in \mathbf{X} is described by:

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

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). We also define $I_f = \{I_f^1, \dots, I_f^p\}$ the set of the sets of indices of exogenous covariates with

$$\forall j \in I_r, I_f^j = \{i | \alpha_i^j \neq 0\} \quad (3.2)$$

$$\forall j \notin I_r, I_f^j = \emptyset. \quad (3.3)$$

Then we have the explicit structure characterized by $S = \{I_f, I_r, \mathbf{p}_f, p_r\}$ where $p_r = |I_r|$, $\mathbf{p}_f = (p_f^1, \dots, p_f^p)$ is the vector of the number of covariates in each sub-regression and $p_f^j = |I_f^j|$, with $|\cdot|$ the cardinal of an ensemble.

The partition of \mathbf{X} implies the uncrossing rule $\mathbf{X}^{I_r} \cap \mathbf{X}^{I_f}$ *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].

We note α the $(p - p_r) \times p_r$ matrix of the α^j . Here we suppose the ϵ_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 α .

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 $\alpha_3 = (1, 1, 0, 0)'$ and we have $S = (\{\{1, 2\}\}, \{3\}, (2), (1))$

3.2 Graph theory

We can model S by a Directed Acyclic Graph (DAG) whose vertices are the p covariates and arcs are the link between them described by the adjacency matrix \mathbf{G} [Bondy and Murty, 1976]. This adjacency matrix is a binary $p \times p$ matrix with $\mathbf{G}_{i,j} = 1$ if and only if $i \in I_f^j$ that is \mathbf{X}^j is explained by \mathbf{X}^i and can also be seen as $\alpha_i^j \neq 0$.

Graphical representation of S helps to understand it and can be compared to the bayesian network representation. It helps to interpret the structure has also been used to construct the algorithm to find S (chapter 5).

The partition of \mathbf{X} mean that the associated graph is bipartite (vertices follow a partition $(\mathbf{X}^{I_r}, \mathbf{X}^{I_f})$) with arcs only going from \mathbf{X}^{I_f} to \mathbf{X}^{I_r} .

We know ([Biggs, 1993]) as a classical result of graph theory that the power of adjacency matrices give the paths in the graph: $\mathbf{G}_{i,j}^k \neq 0$ means that there is a path of length k going from \mathbf{X}^i to \mathbf{X}^j . Because the graph is bipartite and arcs are only going from \mathbf{X}^{I_f} to \mathbf{X}^{I_r} we can deduce that \mathbf{G} is nilpotent: $\mathbf{G}^2 = 0$. And we have the following result: every binary nilpotent matrix of order 2 can be seen as an adjacency matrix of a structure that respects the uncrossing rule. proof by contradiction: if there exist a path of length 2 between some vertices i and j then $\mathbf{G}_{i,j}^2 \neq 0$ so the matrix is not nilpotent of order 2. We can deduce that the number of feasible structure with p covariates is the number of binary nilpotent matrix of order 2.

We see that \mathbf{G} completely describe S and that the sparse storage of \mathbf{G} gives I_f which is sufficient to obtain S by doing $\forall 1 \leq j \leq p : p_f^j = |I_f^j|$, $I_r = \{j | p_f^j > 0\}$ and $p_r = |I_r|$. This decomposition helps us to enumerate all the feasible structure (and thus all the binary nilpotent matrix of order 2).

We note \mathcal{S}_p the set of the feasible structure with p covariates. If we consider all the structure

with equiprobability:

$$S = \{I_f, I_r, \mathbf{p}_f, p_r\} \quad (3.4)$$

$$P(S|p_r) = P(I_f, \mathbf{p}_f|I_r, p_r)P(I_r|p_r) \quad (3.5)$$

$$= P(\mathbf{p}_f|I_f, I_r, p_r)P(I_f|I_r, p_r)P(I_r|p_r) \quad (3.6)$$

$$= P(I_f|I_r, p_r)P(I_r|p_r) \quad (3.7)$$

$$P(I_r|p_r) = \frac{1}{\binom{p}{p_r}} \quad (3.8)$$

$$P(I_f|I_r, p_r) = \frac{1}{(2^{p-p_r} - 1)^{p_r}} \quad (3.9)$$

$$|\mathcal{S}_p| = \sum_{p_r=0}^{p-1} |\mathcal{S}_{p|p_r}| = \sum_{p_r=0}^{p-1} \frac{1}{P(S|p_r)} = \sum_{p_r=0}^{p-1} \binom{p}{p_r} (2^{p-p_r} - 1)^{p_r} \quad (3.10)$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (3.11)$$

is the binomial coefficient.

We have then $|\mathcal{S}_2| = 3, |\mathcal{S}_3| = 13$ and $|\mathcal{S}_{10}| > 13.26 \times 10^9$ so the number of feasible structures really explodes when p is growing.

In the running example: $|\mathcal{S}_5| = 841$

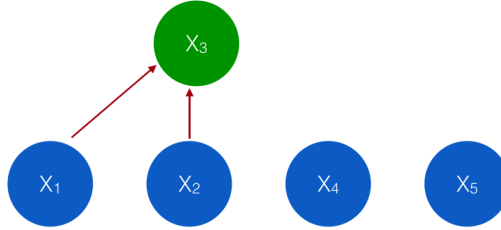


Figure 3.1: The bipartite graph associated to the running example

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

3.3 A by-product model: marginal regression with decor-related covariates

Now we know $P(\mathbf{X}^{I_r}|\mathbf{X}^{I_f}, S)$ by the structure of sub-regressions, we are able to define a marginal regression model $P(\mathbf{Y}|\mathbf{X}^{I_f}, S)$ based on the reduced set of independent covariates $\hat{\beta}_f$ 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 $\mathbf{X} = [\mathbf{X}^{I_f}, \mathbf{X}^{I_r}]$ we can rewrite (2.1):

$$\mathbf{Y}_{|\mathbf{X}^{I_f}, \mathbf{X}^{I_r}, S} = \mathbf{X}^{I_f} \beta_{I_f} + \mathbf{X}^{I_r} \beta_{I_r} + \epsilon_Y \quad (3.12)$$

where $\beta = (\beta_{I_f}, \beta_{I_r}) \in \mathcal{R}^p$ is the vector of the regression coefficients associated respectively to \mathbf{X}^{I_f} and \mathbf{I}_n the identity matrix. We note that (3.1) and (3.12) give also by simple integration on \mathbf{X}^{I_r} a marginal regression model on \mathbf{Y} depending only on uncorrelated covariates \mathbf{X}^{I_f} :

$$P(\mathbf{Y}|\mathbf{X}^{I_f}) = \int_{\mathbf{X}^{I_r}} P(\mathbf{Y}|\mathbf{X}^{I_r}, \mathbf{X}^{I_f}) P(\mathbf{X}^{I_r}|\mathbf{X}^{I_f}) d\mathbf{X} \quad (3.13)$$

$$\mathbf{Y}_{|\mathbf{X}^{I_f}, S} = \mathbf{X}^{I_f}(\beta_{I_f} + \sum_{j \in I_r} \beta_j \alpha_j) + \sum_{j \in I_r} \beta_j \epsilon_j + \epsilon_Y \quad (3.14)$$

$$= \mathbf{X}^{I_f} \beta_{I_f}^* + \epsilon_Y^* \quad (3.15)$$

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

$$\mathbf{Y}_{|\mathbf{X}, S} = \mathbf{X} \beta^* + \epsilon_Y^* \text{ where } \beta^* = (\beta_{I_f}^*, \beta_{I_r}^*) \text{ and } \beta_{I_r}^* = \mathbf{0} \quad (3.16)$$

for which OLS estimator of the coefficients may be biased.

Running example: $\mathbf{Y}_{|\mathbf{X}^{I_f}} = 2\mathbf{x}_1 + 2\mathbf{x}_2 + \mathbf{x}_4 + \mathbf{x}_5 + \epsilon_3 + \epsilon_Y$

3.4 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 $\beta_{I_f}^*$, even with variable selection methods like LASSO or a best subset algorithm like stepwise [Seber and Lee, 2012]. However, we always have $\mathbf{X}^{I_r} = \mathbf{0}$

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 $\hat{\beta}_{I_f}^*$ (irrelevant covariates) and for $\hat{\beta}_{I_r}^* = 0$ (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 \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.

Chapter 4

Numerical results with a known structure

Abstract: Premiers résultats numériques pour une structure (hors coefficients de sous-régression) connue. On constate un net apport de la méthode de préselection.

4.1 Illustration of the tradeoff conveyed by the pre-treatment

We compare the OLS estimator on \mathbf{X} defined in section 2.2 with the estimator obtained by the pre-treatment that is \mathbf{X}^{I_f} selection.

For the marginal regression model defined in (3.15) we have the OLS unbiased estimator of β^* :

$$\hat{\beta}_{I_f}^* = (\mathbf{X}^{I_f'} \mathbf{X}^{I_f})^{-1} \mathbf{X}^{I_f'} \mathbf{Y} \text{ and } \hat{\beta}_{I_r}^* = \mathbf{0} \quad (4.1)$$

We see in (3.14) that it gives an unbiased estimation of \mathbf{Y} and β^* but in terms of β this estimator is biased:

$$\mathbb{E}[\hat{\beta}_{I_f}^* | \mathbf{X}^{I_f}] = \beta_{I_f} + \sum_{j \in I_r} \beta_j \alpha_j \text{ and } \mathbb{E}[\hat{\beta}_{I_r}^* | \mathbf{X}^{I_f}] = \mathbf{0} \quad (4.2)$$

with variance:

$$\text{Var}[\hat{\beta}_{I_f}^* | \mathbf{X}^{I_f}] = (\sigma_Y^2 + \sum_{j \in I_r} \sigma_j^2 \beta_j^2) (\mathbf{X}^{I_f'} \mathbf{X}^{I_f})^{-1} \text{ and } \text{Var}[\hat{\beta}_{I_r}^* | \mathbf{X}^{I_f}] = \mathbf{0} \quad (4.3)$$

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

$$E[\text{MSE}(\hat{\beta} | \mathbf{X})] = \|\text{Bias}\|_2^2 + \text{Tr}(\text{Var}(\hat{\beta})) \quad (4.4)$$

$$E[\text{MSE}(\hat{\beta}_{OLS} | \mathbf{X})] = 0 + \sigma_Y^2 \text{Tr}((\mathbf{X}' \mathbf{X})^{-1}) \quad (4.5)$$

$$E[\text{MSE}(\hat{\beta}_{OLS}^* | \mathbf{X})] = \left\| \sum_{j \in I_r} \beta_j \alpha_j \right\|_2^2 + \|\beta_{I_r}\|_2^2 + (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 \beta_j^2) \text{Tr}((\mathbf{X}^{I_f'} \mathbf{X}^{I_f})^{-1}) \quad (4.6)$$

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 4.1 shows the theoretical MSE evolution with the strength of the sub-regression. In this section, all experiences have been made 100 times to obtain smooth curves. So we have generated 100 times \mathbf{X} and \mathbf{Y}

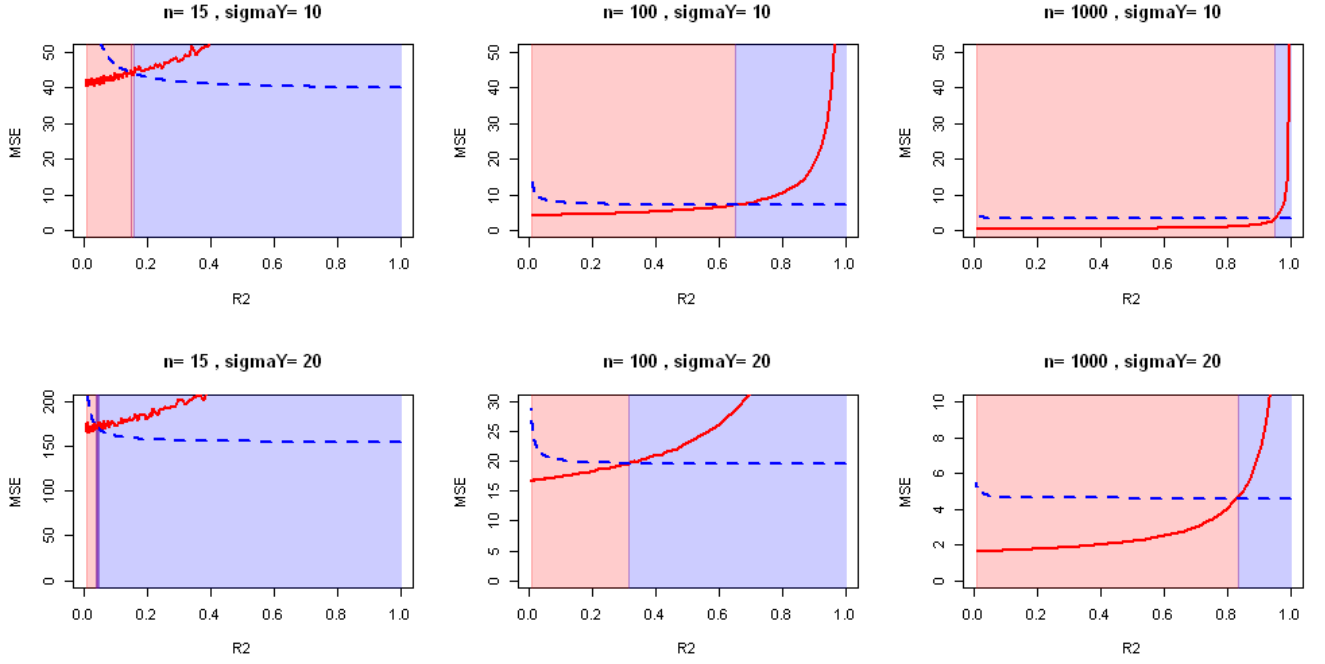


Figure 4.1: Theoretical MSE on β of OLS (red) and CorReg's marginal model(blue) estimators for varying R^2 of the sub-regression, n and σ_Y .

It is clear in Figure 4.1 that the marginal model is more robust than OLS on \mathbf{X} . And when sub-regression get weaker (R^2 tends to 0) 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 $\text{Tr}(\text{Var}(\hat{\beta})) = \text{Tr}(\text{Var}(\hat{\beta}_{I_f})) + \text{Tr}(\text{Var}(\hat{\beta}_{I_r}))$ for both models but for the marginal model $\text{Tr}(\text{Var}(\hat{\beta}_{I_r})) = 0$. 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).

But it is only the theoretical MSE and we want to know what happens in the real life. So we look at the empirical MSE on both $\hat{\beta}$ and \hat{Y} . Here again, each configuration is computed 100 times and we take the mean to smooth the curves. We first observe that real results are better from expected (Figure ??). This comes from usage of QR decomposition to inverse matrices, that are less impacted by ill-conditioned matrices. Our package CorReg also uses this decomposition. But our marginal model remains better for strong correlations.

Further results are provided in sections 6 and 7.

4.2 Predictive efficiency

Résultats comme dans l'article mais sur vrai S modèles avec beaucoup de variables (car rapide vu que pas de MCMC). Faire avec vrai modèle= modèle complet puis X1 uniquement puis X2 uniquement puis aussi Amax=15 si p largement supérieur à 100

Chapter 5

Estimation of the Structure of subregression by MCMC

Abstract: La structure de sous-régression est supposée inconnue. Il nous faut donc la trouver. Un algorithme de type MCMC est proposé pour résoudre cette problématique. La mise en oeuvre de celui-ci passe par une modélisation complète du jeu de données et nous pousse à introduire un nouveau critère de choix de modèle tenant compte du nombre de modèles testés.

5.1 Bayesian approach

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 (see section 5.8). 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 \mathbf{Y} because the structure is inherent to \mathbf{X} only. Thus it must be a global criterion. Because it is about model selection, 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 $P(S|\mathbf{X})$ and we have:

$$P(S|\mathbf{X}) \propto P(\mathbf{X}|S)P(S) = P(\mathbf{X}^{I_r}|\mathbf{X}^{I_f}, S)P(\mathbf{X}^{I_f}|S)P(S) \quad (5.1)$$

So we will try to maximize $\psi(\mathbf{X}, S) = P(\mathbf{X}|S)P(S)$. It will be done with a Markov chain Monte-Carlo algorithm (MCMC).

5.2 Sub-regression structure in details

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

To be able to compare structures with $P(S|\mathbf{X})$, we need a full generative model on \mathbf{X} . Sub-regressions give $P(\mathbf{X}^{I_r}|\mathbf{X}^{I_f}, S)$ but $P(\mathbf{X}^{I_f}|S)$ is still undefined. We suppose that variables in

\mathbf{X}^{I_f} follow Gaussian mixtures of $K_j > 0$ components:

$$\forall \mathbf{X}^j \notin \mathbf{X}^{I_r} : \mathbf{X}_{|S}^j \sim f(\boldsymbol{\theta}_j) = \mathcal{GM}(\boldsymbol{\pi}_j; \boldsymbol{\mu}_j; \boldsymbol{\sigma}_j^2) \text{ with } \boldsymbol{\pi}_j, \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2 \text{ vectors of size } K_j. \quad (5.2)$$

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.

We now have a full generative model on \mathbf{X} .

5.2.2 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 that we 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 ($\exists j, \sigma_j^2 = 0$), 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 $\forall j, \sigma_j^2 \neq 0$, then $\mathbf{X}^{I_f'} \mathbf{X}^{I_f}$ and $\mathbf{X}' \mathbf{X}$ 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: $\mathbf{X}^{I_r} = \mathbf{y}^{S^c}$ and $\mathbf{X}^{I_f} = \mathbf{y}^R$. We have supposed independence between variables in \mathbf{X}^{I_f} so the identifiability theorem from Maugis tells that our model is identifiable if variables in \mathbf{X}^{I_f} are Gaussian mixtures.

We define $\mathbf{X}^G \subsetneq \mathbf{X}^{I_f}$ containing Gaussian variables and we note the Gaussian mixtures $\mathbf{X}^{G^c} \neq \emptyset$ its complement in \mathbf{X}^{I_f} . We suppose that variables in \mathbf{X}^{I_r} are all Gaussian mixtures. It implies that $\forall j \in I_r, \exists i \in I_f^j$ so that $\mathbf{X}^i \subset \mathbf{X}^{G^c}$ 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 $\mathbf{X}^{I_r} = \mathbf{X}^{I_f} \boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ 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.

$$\mathbf{X}_{r|\mathbf{X}^G, \mathbf{X}^{G^c}} = \mathbf{X}^G \boldsymbol{\alpha}_G + \mathbf{X}^{G^c} \boldsymbol{\alpha}_{G^c} + \boldsymbol{\varepsilon} \quad (5.3)$$

$$\mathbf{X}_{r|\mathbf{X}^{G^c}} = \mathbf{X}^{G^c} \boldsymbol{\alpha}_{G^c} + \tilde{\boldsymbol{\varepsilon}} \text{ is identifiable where} \quad (5.4)$$

$$\tilde{\boldsymbol{\varepsilon}}_j = \mathbf{X}^G \boldsymbol{\alpha}_j^G + \boldsymbol{\varepsilon}_j \text{ is Gaussian.} \quad (5.5)$$

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

Remark: Identifiability of S is not necessary to use a given structure but helps to find it. In the followings, true S is supposed to be identifiable (at least one Gaussian mixture in each sub-regression).

5.3 Sub-regression model selection

5.3.1 Bayesian criterion for quality

Our full 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 \mathbf{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*, that is more classical.

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

When considering (5.1) we see that uniform law on $P(S)$ gives $\psi(\mathbf{X}, S) \propto P(\mathbf{X}|S)$ so it is equivalent to a minimization of the *BIC*. We note Θ the set of the parameters of the generative model

$$-2 \log P(\mathbf{X}|S) \approx BIC = -2\mathcal{L}(\mathbf{X}, S, \Theta) + |\Theta| \log(n) \quad (5.6)$$

But *BIC* tends to give too complex structures because we test a great range of models and the number of model compared is not taken into account [Massart and Picard, 2007]. Thus we choose to penalise the complexity a bit more. We don't want to modify the *BIC* to keep its properties.

We have the explicit structure characterized by $S = \{I_f, I_r, p_f, p_r\}$, then we suppose a hierarchical uniform *a priori* distribution $P(S) = P(I_f|\mathbf{p}_f, I_r, p_r)P(\mathbf{p}_f|I_r, p_r)P(I_r|p_r)P(p_r)$ instead of the simple uniform law on S that is generally used and provides no penalty. It goes against the fact that the number of models with p_r sub-regressions and the number of possible combination for each sub-regression depends on p_r and thus provides distinct penalties according to the complexity. Thus we have :

$$BIC_+(X|S) = BIC(X|S) - \ln(P(S)) \quad (5.7)$$

The hierarchical uniform hypothesis gives:

$$P(S) = P(I_f|\mathbf{p}_f, I_r, p_r)P(\mathbf{p}_f|I_r, p_r)P(I_r|p_r)P(p_r) \text{ with} \quad (5.8)$$

$$P(p_r) = \frac{1}{p} \quad (5.9)$$

$$P(I_r|p_r) = \frac{1}{\binom{p}{p_r}} \quad (5.10)$$

$$P(\mathbf{p}_f|I_r, p_r) = \frac{1}{p_r \times \frac{1}{p-p_r}} = \frac{p-p_r}{p_r} \quad (5.11)$$

$$P(I_f|\mathbf{p}_f, I_r, p_r) = \frac{1}{\prod_{j \in I_r} \binom{p-p_r}{p_f^j}} \quad (5.12)$$

instead of $P(S) = \frac{1}{|S_p|}$ as defined in section 3.2 for the classical uniform hypothesis. It increases penalty on complexity for $p_r \leq \frac{p}{2}$ and $p_f^j \leq \frac{p}{2}$ because probability of a complex model is under-estimated. 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 6 and 7). 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.

5.3.3 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 (5.1)) 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 compared to the true model
- $\Delta compl$: the difference in complexity between both model:
 $\Delta compl = \sum_{j \in p_r} p_f^j - \sum_{j \in \hat{p}_r} \hat{p}_f^j$

5.4 Neighbourhood

We note \mathcal{S}_p the ensemble of feasible structures of size p (those uncrossed, *i.e.* with $I_f \cap I_r = \emptyset$). For each step q , starting from $S \in \mathcal{S}_p$ we define a neighbourhood:

$$\mathcal{V}_S = \{S\} \cup \{S^{(i,j)} | (i,j) \in \mathcal{A}_q\} \cap \mathcal{S}_p \quad (5.13)$$

where \mathcal{A}_q is a set of arcs to modify (add or remove) in the associated graph, defined according to a strategy. And we have for given S and (i,j) :

- if $i \in I_f^j$ then we remove i from I_f^j (arc removal)
- else we had i in I_f^j

Then coherence between I_f and others parts of S is done by $\forall 1 \leq j \leq p : p_f^j = |I_f^j|$, $I_r = \{j | p_f^j > 0\}$ and $p_r = |I_r|$.

In the adjacency matrix we just do: $G_{i,j} = 1 - G_{i,j}$. The main advantage of such a neighbourhood is that increasing and decreasing complexities are tested at each step without arbitrary ratio. If we just look at the sub-regression system, we have to choose for each sub-regression if we add, remove or keep covariates and we also have to choose if we had or delete some sub-regression. Adjacency matrix makes the neighbourhood extremely natural with just the modification of a value in a binary matrix.

5.4.1 Strategy

Many strategies can be imagined. First, we can decide to keep the local \hat{S} in the neighbourhood or not, that is allowing or not stationarity. Here the MCMC is not used for sampling or density estimation. We just want to find the structure with the best value of $\psi(\mathbf{X}, S)$ so it is not an evidence to allow or not stationarity. Our package CorReg give the user the choice with stationarity, included in the neighbourhood by default.

The only constraint on \mathcal{A}_q is that $\forall (i, j) \in \mathcal{A}_q, i \neq j$
We propose, for step q to draw j from $\mathcal{U}(\{1, \dots, p\})$ and then

$$\mathcal{A}_{q|j} = \{(i, j) | i \neq j\} \quad (5.14)$$

Such a strategy can be interpreted as the uniform choice of a sub-regression to modify followed by the proposal of each possible unary change. Our package CorReg let the user choose many other strategies like a fixed number of random couples (i, j) , or the union of the j^{th} line and column of G .

5.4.2 Active relaxation of the constraints

We have defined the neighbourhood with an intersection with \mathcal{S}_p . In practice, for some of the $(i, j) \in \mathcal{A}_q$, we have $S^{(i,j)} \notin \mathcal{S}_p$. Such candidates are basically rejected so the number of candidates is not constant at each step. Moreover, complex structures reduce the size of the potential neighbourhood because of the uncrossing rule. Thus we propose a relaxation method by a new definition of $S^{(i,j)}$:

- if $i \notin I_f^j$ (add):
 - $I_f^j = I_f^j \cup \{i\}$
 - $I_f^i = \emptyset$ (explicative variables can't depend on others : column-wise relaxation)
 - $I_f = I_f \setminus \{j\}$ (dependent variables can't explain others : row-wise relaxation)
- else (remove): $I_f^j = I_f^j \setminus \{i\}$

It can be seen as forcing the modification by deleting what would make the structure not feasible. So in one step we can test a model that remove completely a sub-regression, remove the explicative role of a covariate in all sub-regression and create a new pairwise sub-regression. It drastically increases the scope of the neighbourhood and guarantee to always have the same number of candidates during the MCMC. It can be compared to simulated annealing that sometimes proposes exotic candidates to avoid local extrema, but here without any temperature to set. Here again, the neighbourhood remains natural, without arbitrary parameters to tune. Another advantage of the relaxation method is that it reduce complexity very quickly without having to deconstruct a sub-regression, so it helps to have simpler models in a small amount of time (asymptotical results are the same because the chain is regular thus ergodic).

5.5 The walk

Once we have a neighbourhood, we have to choose a candidate for the next step. The walk follows a time-homogeneous Markov Chain whose transition matrix \mathcal{P} has $|\mathcal{S}_p|$ rows and columns (combinatory so we just compute the probabilities when we need them). At each step the Markov chain moves with probability:

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) = \sum_{j=1}^p \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(-\frac{1}{2}\psi(\mathbf{X}, \tilde{S}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(-\frac{1}{2}\psi(\mathbf{X}, S_l))} \quad (5.15)$$

And \mathcal{S}_p is a finite state space.

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [Grinstead and Snell, 1997] and the output will be the best structure in terms of $P(S|\mathbf{X})$ which weights each candidate. Practically speaking,

CORREG returns the best structure seen during the walk (even if the corresponding candidate has never been chosen). The package also give the local structure when the walk stops so user can relaunch the algorithm from the same point if he wants to go further. The main criterion to stop the walk is a maximum number of iteration but CorReg can also stop the walk after a given number of step on the best model found. Numerical results (Section 4) illustrates the efficiency of the walk when the true model contains structures with various strength (Figure ??) and an example with a non-linear structure (Figure ??).

5.6 Initialization

5.6.1 Correlation-based initialization

If we have some knowledge about some sub-regressions (physical models for example) we can add them in the found and/or initial structure. So the model is really expert-friendly. The initial structure can be based on a first warming algorithm taking the correlations into account. Coefficients are randomly placed into I_f , according to a Bernoulli draw weighted by the absolute value of the correlations and with respect to the uncrossing constraint. Uncrossing constraint will not allow some strong correlation to be taken into account according to the order of the Bernoulli drawing so we can draw with a random order or by ordering by descending correlations.

We note that the *BIC* associated to initial model is often better than the *BIC* of the void structure, so we compare several chains in Figures 5.1 and 5.2:

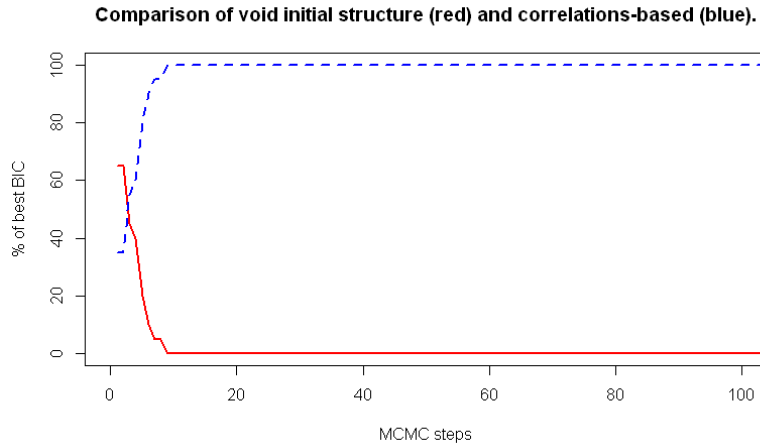


Figure 5.1: Amount of time each method is better for the 100 first steps of the MCMC.

5.6.2 Multiple intialisation

Local extremas are a known issue for most of optimization methods, and one would rather test multiple short chains than lose time in initialisation or long chains [Gilks et al., 1996]. We also compare the results obtained with several number of chains. Figure 5.3 shows the evolution of the *BIC* of the best chain with a number of chains varying from 1 to 10, so the model with 10 chains contain the others and is almost as good as they are. We see that multiple

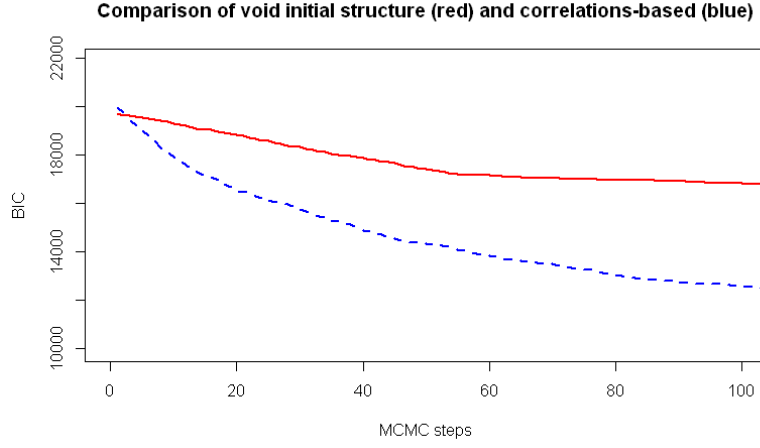


Figure 5.2: Evolution of the BIC (criterion to minimize in the MCMC) for each method.

initialization is efficient but the gain is logarithmic in the number of tries so it is recommended to use multiple chains but not too much.

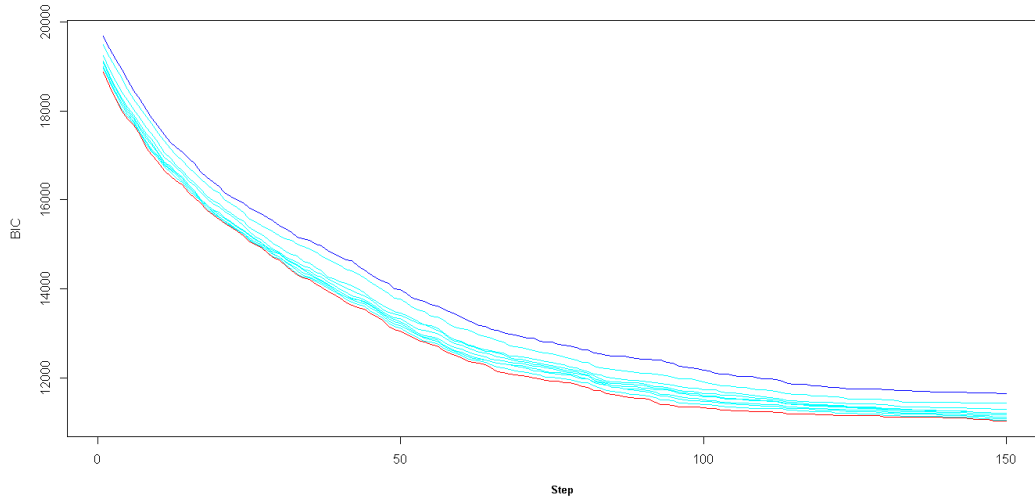


Figure 5.3: Comparison of distinct number of correlation-based initialisations for the MCMC.

5.7 Pruning

If the complexity of S remains too high, pruning methods can be used.

Variable selection

We can use variable selection methods like the LASSO to estimate the coefficient α and obtain some supplementary zeros. The complexity may be the same so it is not really a pruning method.

R^2 thresholding

We can define a minimal value for the R^2 of the sub-regression to maintain them in the final structure. But this minimal value would be totally arbitrary and we know that it is frequent to use linear regression with real datasets that only show a R^2 between 0.1 and 0.2. It is particularly true in social sciences.

Test

Another pruning method would be to delete sub-regression that offer a F-statistic under a minimal value. In the followings, the chain was launched with twenty initialisations each time, based on the correlation matrix.

Additional cleaning steps

Because the walk is not exhaustive, it does make sense to let the walk continue a few steps with neighbourhood containing only suppressions in the structure. The walk follows an ergodic chain thus it is just an heuristic change in the strategy with:

$$\mathcal{A}_q = \{(i, j) | i \in I_f^j\} \quad (5.16)$$

It is not based on any arbitrary parameter and change the result only if it finds a better structure in terms of the criterion ψ used in the walk.

5.8 The Graphical LASSO

Graphical LASSO [Friedman et al., 2008] [Witten et al., 2011] [Tibshirani et al.,] [Friedman et al., 2010] is set to give undirectionnal (thus symmetric) graphs by selection in the precision matrix (the inverse of the variance-covariance matrix). It does make sense for exponential family because in these cases, zeros in the precision matrix Σ^{-1} can be interpreted in terms of conditionnal independence between covariates [Dempster, 1972]. But we have supposed Gaussian mixture on \mathbf{X} and we search an oriented graph. But we can still use it for initialization, for example by a hadamard product with G_0 the adjacency matrix of the initial structure. We can also try to give the graph a bipartite orientation. We first have to obtain a bipartite graph, that mean to have no even cycles. A particular case would be the minimum spanning tree [Graham and Hell, 1985, Moret and Shapiro, 1991, Gower and Ross, 1969]. But it is time consuming and has no theoretical properties relied to our problematic of minimizing ψ , so the idea was left behind after some tries.

5.9 CorReg

The CorReg package is now on CRAN and provides many parameters for the walk. If wanted it can return some curves associated to the walk to have an idea of what happens with distinct strategies.

We define the complexity of a structure S as the number of elements in the adjacency matrix, that is the number of links between covariates and is obtained by:

$$\text{Complexity}(S) = \sum_{j \in I_r} p_f^j \quad (5.17)$$

We compare some walks with each time the same dataset and the same seed for the random generator. We have $p = 100$ and $n = 50$.

For Figures 5.4 and 5.5 we start from an arbitrary structure with a complexity of 62. We see that relaxation helps to delete these false sub-regressions and avoid to be stuck in it, improving the BIC much faster. We also observe that final complexities are comparable. Here the MCMC was launched only once (with the totally arbitrary initial structure based on nothing), the true structure had a complexity of 120.

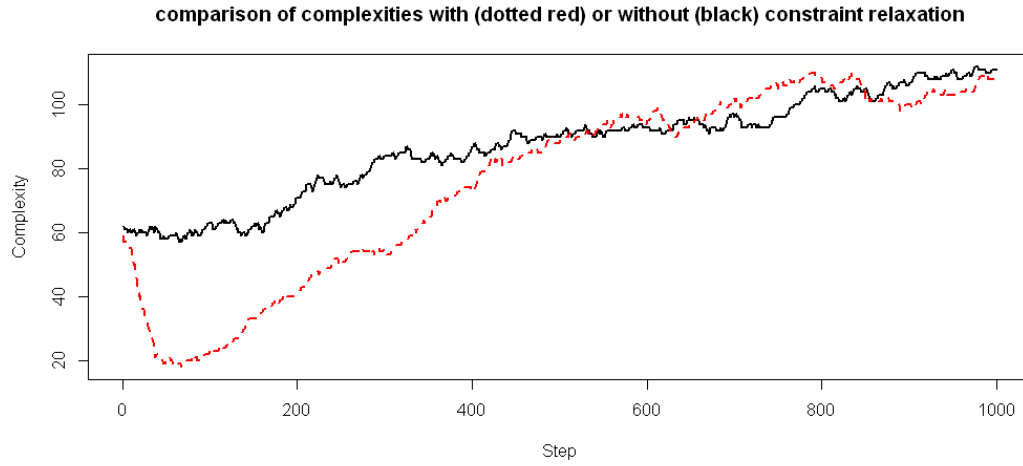


Figure 5.4: Comparison of complexity evolution with or without constraint relaxation.

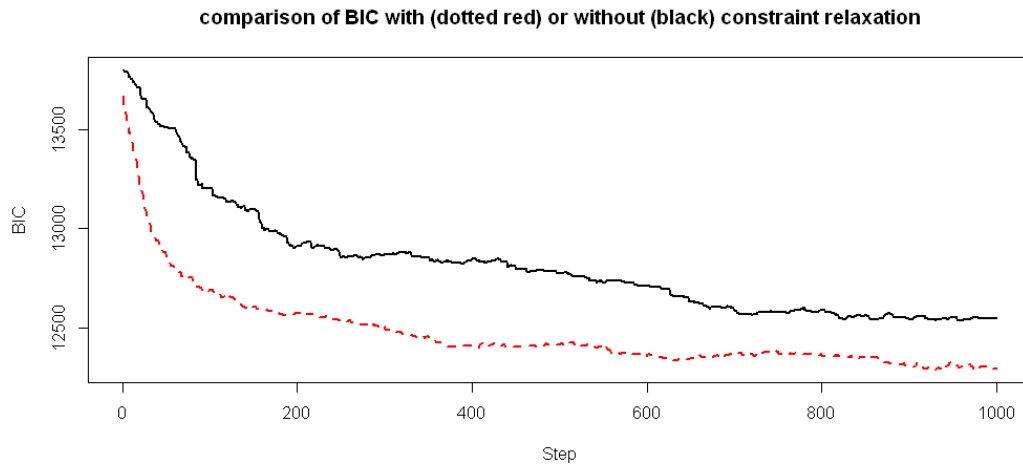


Figure 5.5: Comparison of BIC evolution with or without constraint relaxation.

Chapter 6

Numerical results on simulated datasets

6.1 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 6.2 shows the results obtained in terms of \hat{S} . Sections ?? and 6.3.3 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 \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.

6.2 Results on \hat{S}

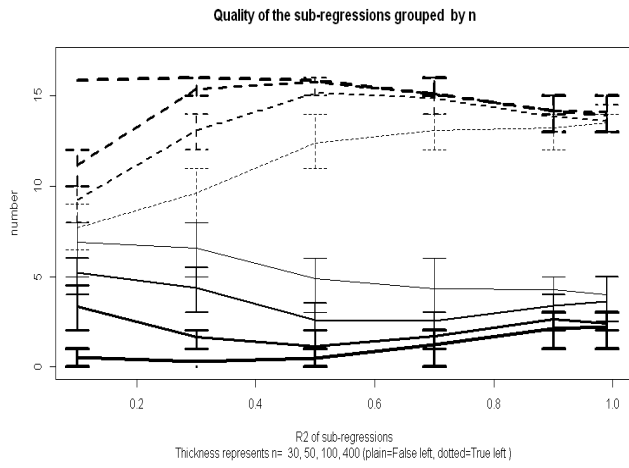


Figure 6.1: Quality of the subregressions found with classical BIC criterion

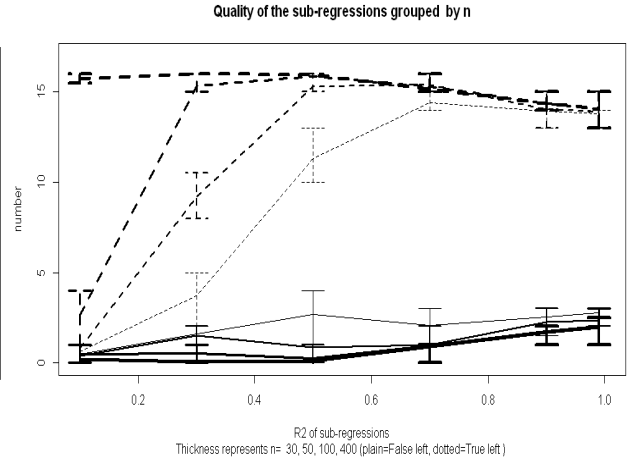


Figure 6.2: Quality of the subregressions found with our BIC_+ criterion

6.3 Results on prediction

6.3.1 Y depends only on covariates in X^{I_f} (best case for us)

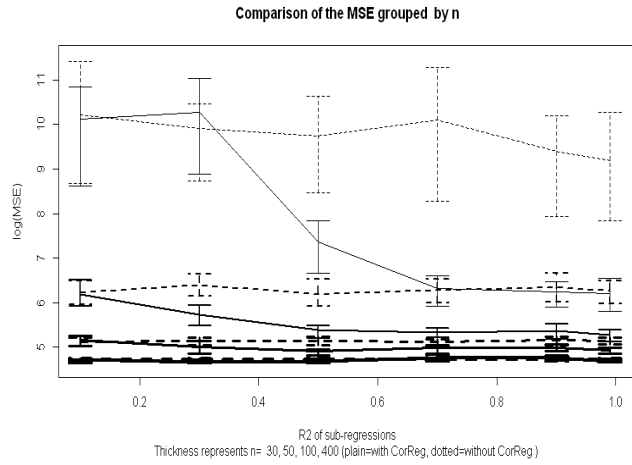


Figure 6.3: Comparison of the MSE between OLS and CorReg+OLS

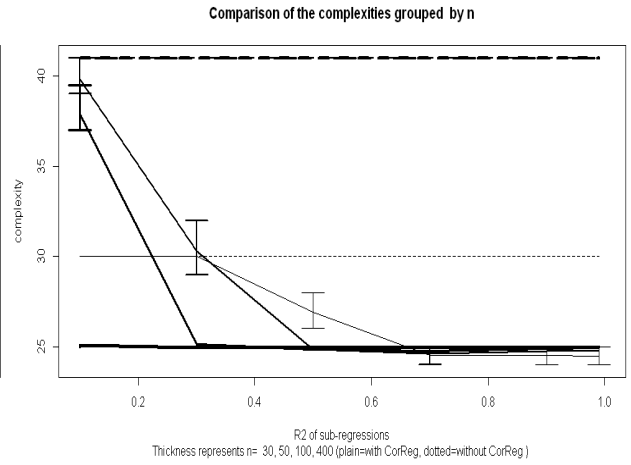


Figure 6.4: Comparison of the complexities between OLS and CorReg+OLS

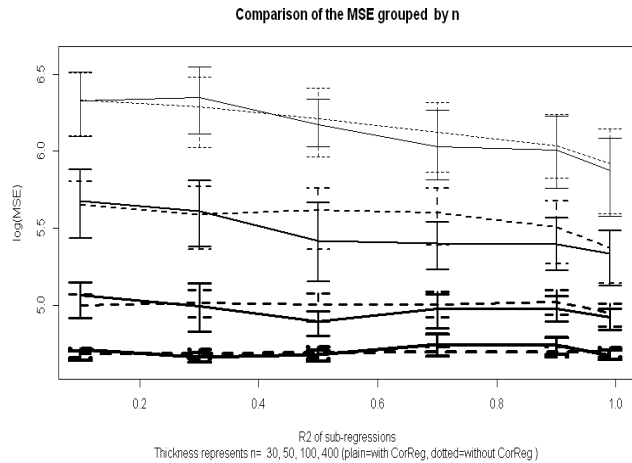


Figure 6.5: Comparison of the MSE between LASSO and CorReg+LASSO

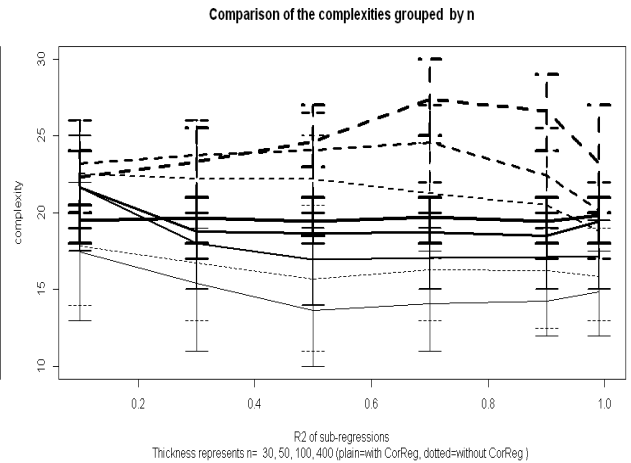


Figure 6.6: Comparison of the complexities between LASSO and CorReg+LASSO

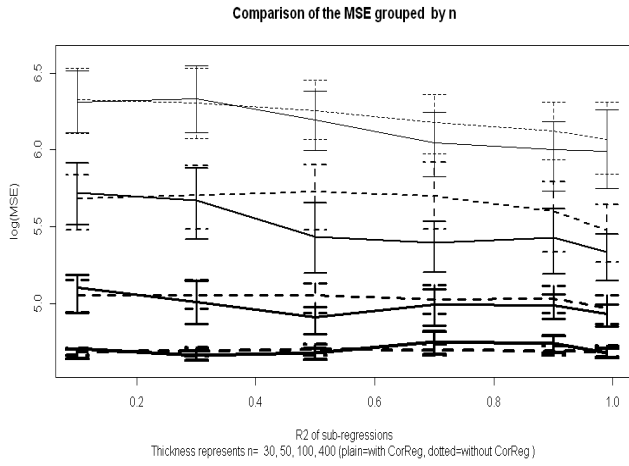


Figure 6.7: Comparison of the MSE between elasticnet and CorReg+elasticnet

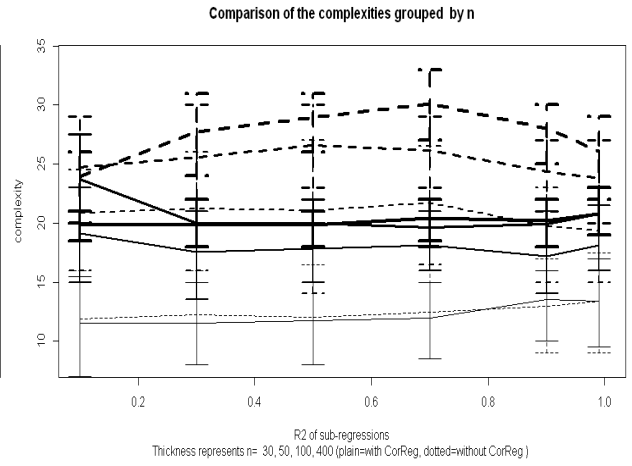


Figure 6.8: Comparison of the complexities between elasticnet and CorReg+elasticnet

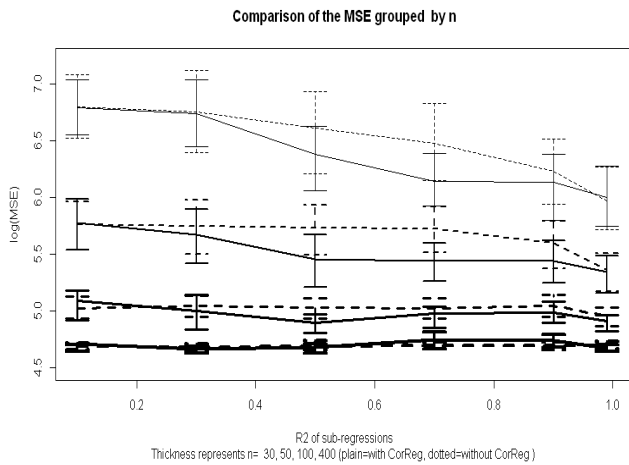


Figure 6.9: Comparison of the MSE between stepwise and CorReg+stepwise

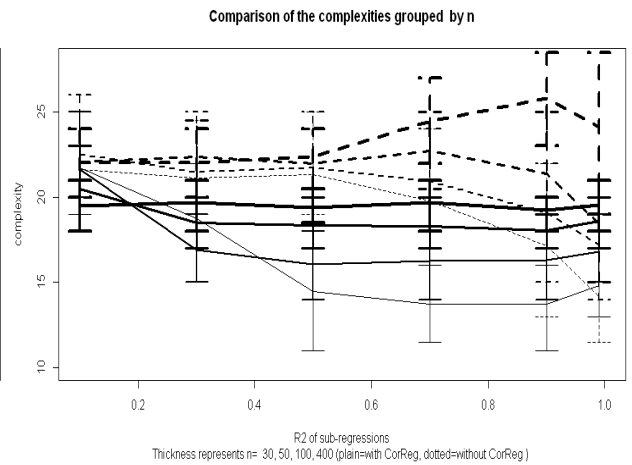


Figure 6.10: Comparison of the complexities between stepwise and CorReg+stepwise

6.3.2 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).

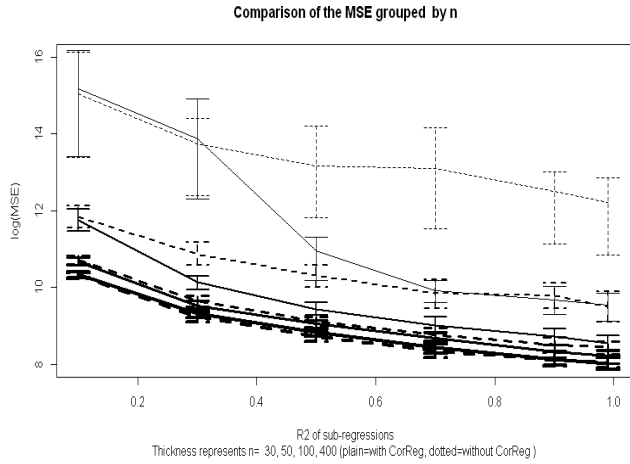


Figure 6.11: Comparison of the MSE between OLS and CorReg+OLS

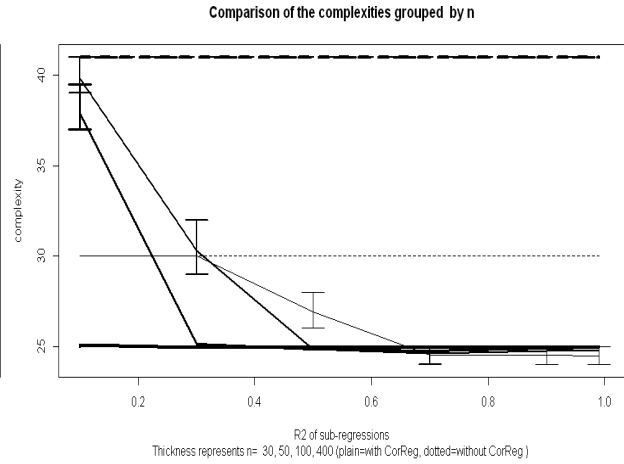


Figure 6.12: Comparison of the complexities between OLS and CorReg+OLS

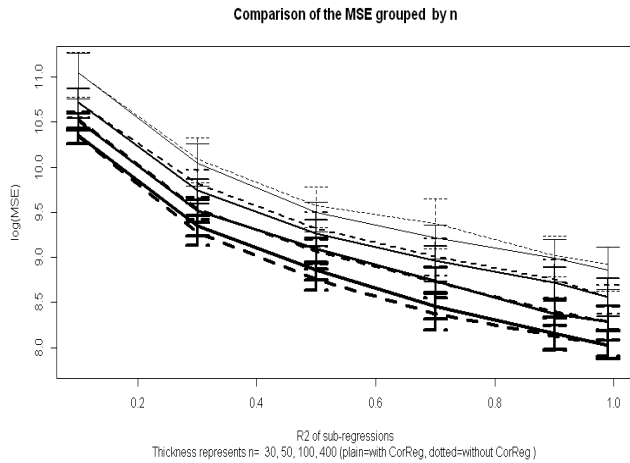


Figure 6.13: Comparison of the MSE between LASSO and CorReg+LASSO

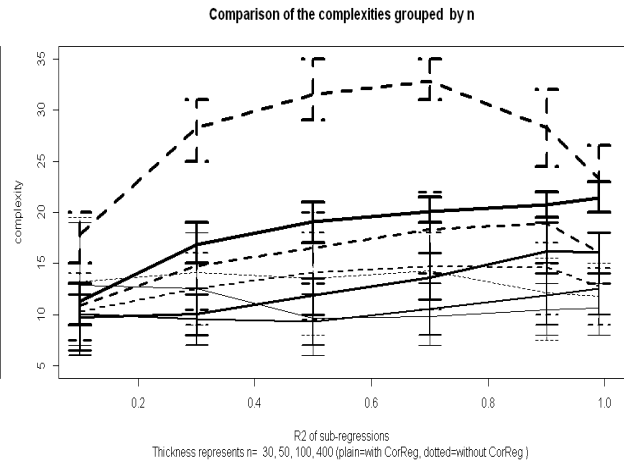


Figure 6.14: Comparison of the complexities between LASSO and CorReg+LASSO

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).

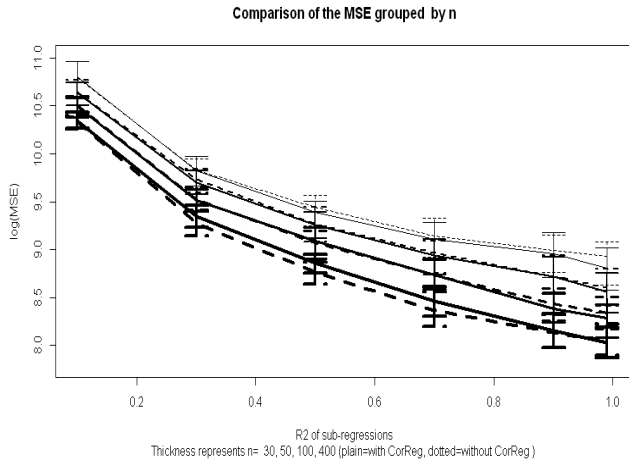


Figure 6.15: Comparison of the MSE between elasticnet and CorReg+elasticnet

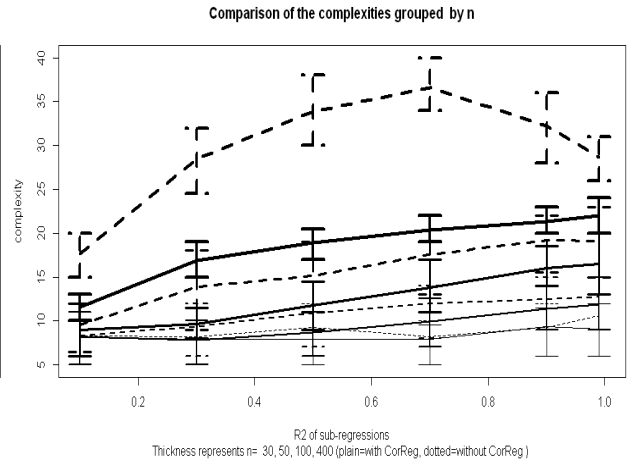


Figure 6.16: Comparison of the complexities between elasticnet and CorReg+elasticnet

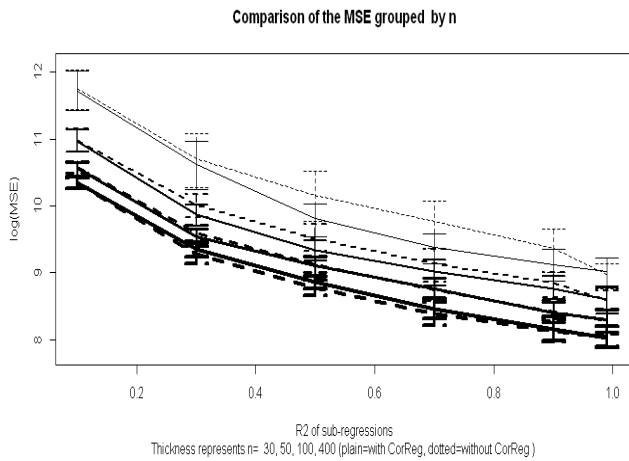


Figure 6.17: Comparison of the MSE between stepwise and CorReg+stepwise

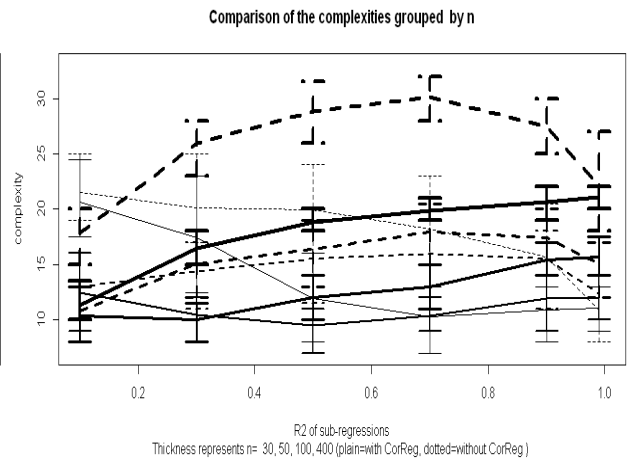


Figure 6.18: Comparison of the complexities between stepwise and CorReg+stepwise

6.3.3 Y depends only on covariates in \mathbf{X}^{I_r} (worst case for us)

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

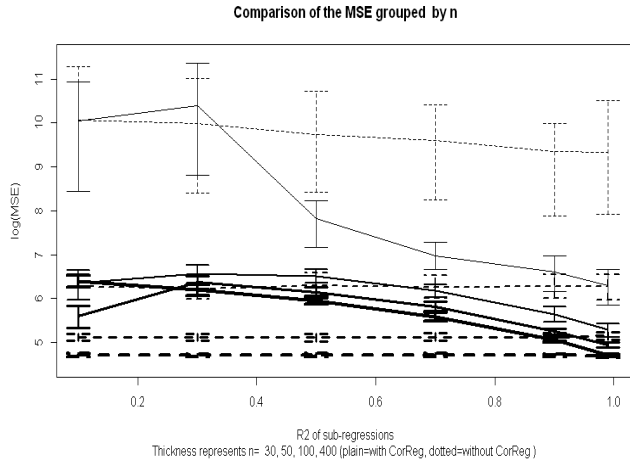


Figure 6.19: Comparison of the MSE between OLS and CorReg+OLS

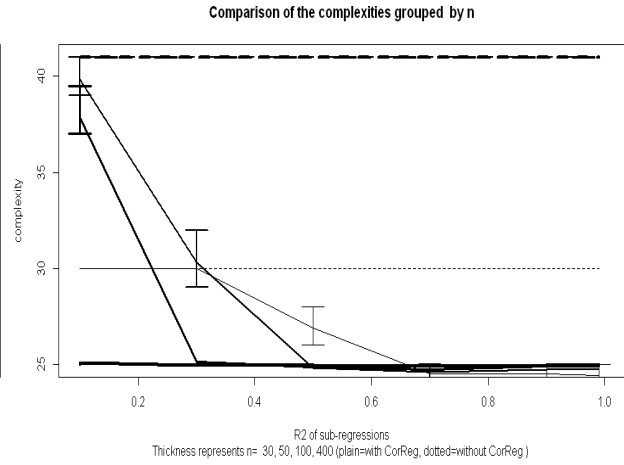


Figure 6.20: Comparison of the complexities between OLS and CorReg+OLS

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

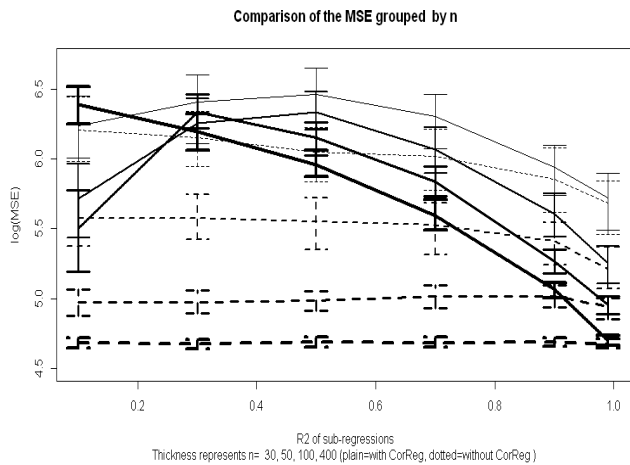


Figure 6.21: Comparison of the MSE between LASSO and CorReg+LASSO

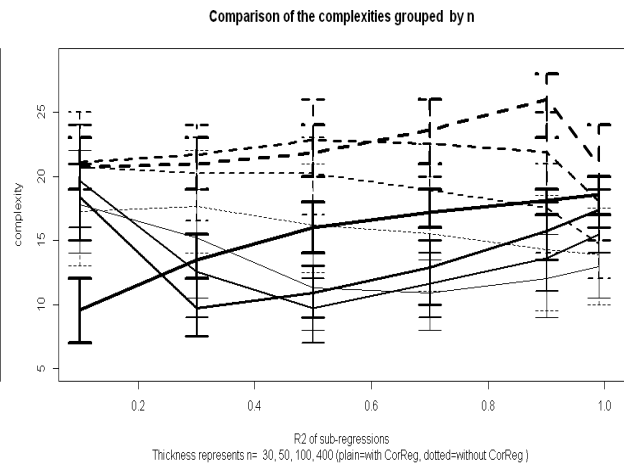


Figure 6.22: Comparison of the complexities between LASSO and CorReg+LASSO

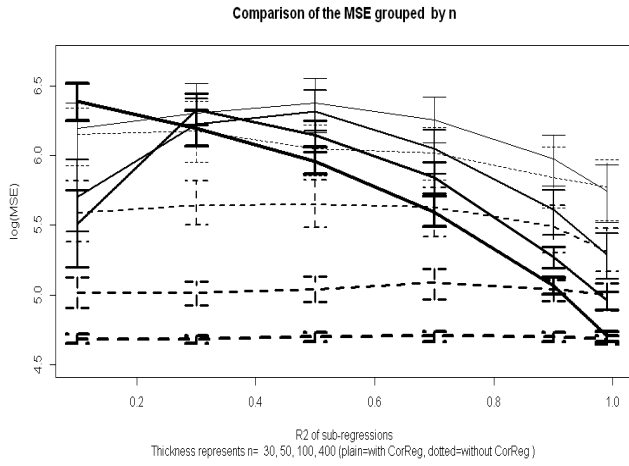


Figure 6.23: Comparison of the MSE between elasticnet and CorReg+elasticnet

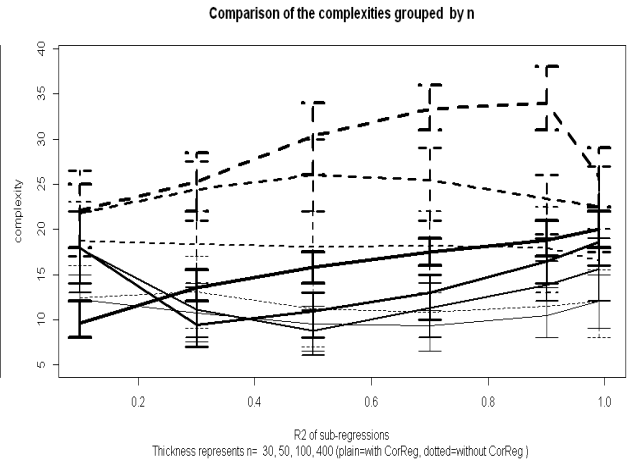


Figure 6.24: Comparison of the complexities between elasticnet and CorReg+elasticnet

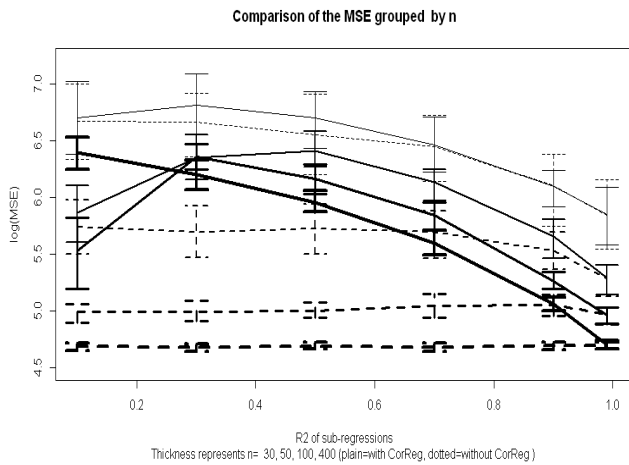


Figure 6.25: Comparison of the MSE between stepwise and CorReg+stepwise

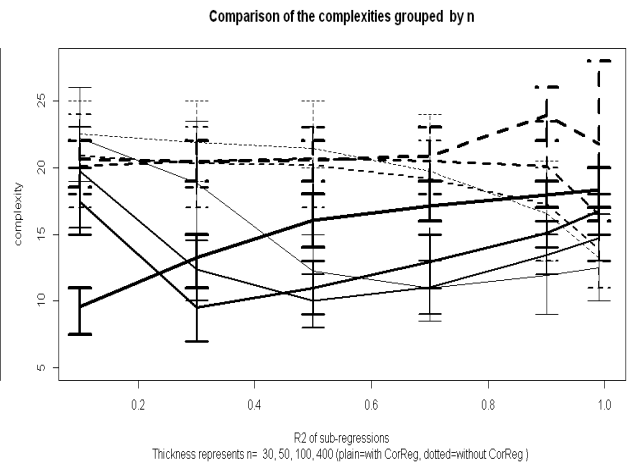


Figure 6.26: Comparison of the complexities between stepwise and CorReg+stepwise

Chapter 7

Numerical results on real datasets

7.1 Quality case study

This work takes place in steel industry context, with quality oriented objective : to understand and prevent quality problems on finished product, knowing the whole process. The correlations are strong here (many parameters of the whole process without any a priori and highly correlated because of physical laws, process rules, *etc.*).

We have :

- a quality parameter (confidential) as response variable,
- 205 variables from the whole process to explain it.

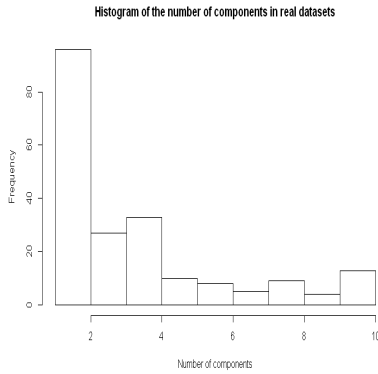


Figure 7.1: Distribution of the number of components found for each covariate.

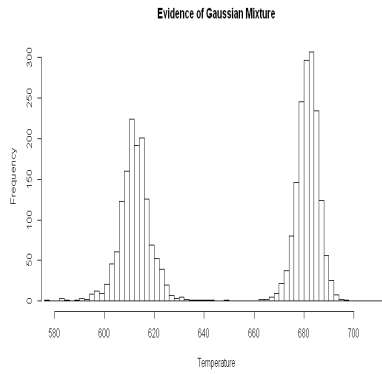


Figure 7.2: Example of non-Gaussian real variable easily modeled by a Gaussian mixture.

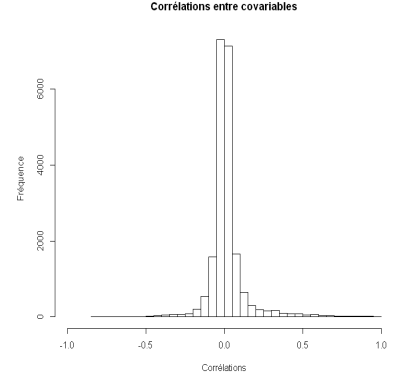


Figure 7.3: Histogram of correlations in \mathbf{X} .

We get a training set of $n = 3000$ products described by $p = 205$ variables from the industrial process and a validation sample of 847 products. Let's note ρ the absolute value of correlations between two covariates. Industrial variables are naturally highly correlated as the width and the weight of a steel slab ($\rho = 0.905$), the temperature before and after some tool ($\rho = 0.983$), the roughness of both faces of the product ($\rho = 0.919$), a mean and a max ($\rho = 0.911$).

The objective here is not only to predict non-quality but to understand and then avoid it. CorReg provides an automatic method without any a priori and is combined with variable selection methods. So it allows to obtain in a small amount of time some indications on the source of the problem, and to use human resources efficiently. When quality crises occurs,

time is extremely precious so automation is a real stake. The combinatorial aspect of the sub-regression models makes it impossible to do manually.

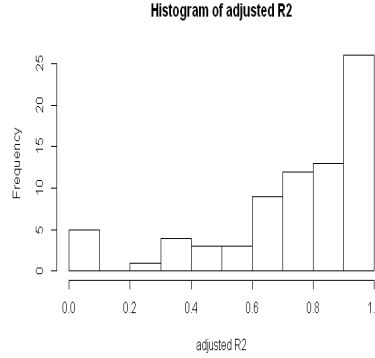


Figure 7.4: R_{adj}^2 of the 76 sub-regressions.

CORREG found the above correlations but it also found more complex structures describing physical models, like $\text{Width} = f(\text{Mean flow}, \text{Mean speed})$ even if the true Physical model is not linear : $\text{Width} = \text{flow} / (\text{speed} * \text{thickness})$ (here thickness is constant). Non-linear regulation models used to optimize the process were also found (but are confidential). These first results are easily understandable and meet metallurgists expertise. Sub-regressions with small values of R^2 are associated with non-linear model (chemical kinetics for example). The algorithm gives a structure of $p_r = 76$ subregressions with a mean of $\bar{p}_f = 5.17$ regressors. In \mathbf{X}^{I_f} the number of $\rho > 0.7$ is **79.33%** smaller than in \mathbf{X} .

It is now time to look at the predictive results (Table 7.1). We see that CorReg improves the results for each method tested in terms of prediction. We get parsimonious models automatically in a small amount of time (several hours but able to work during the night or the week-end)

Method	indicator	With CorReg	without CorReg
OLS	MSE	13.30	14.03
	(complexity)	(130)	(206)
LASSO	MSE	12.77	12.96
	(complexity)	(24)	(21)
Elasticnet	MSE	12.15	13.52
	(complexity)	(40)	(78)
Ridge	MSE	12.69	13.09
	(complexity)	(130)	(206)

Table 7.1: Results obtained on a validation sample (847 individuals).

In terms of interpretation, the main regression comes with the family of regression so it gives a better understanding of the consequences of corrective actions on the whole process. It typically permits to determine the *tuning parameters* whereas variable selection alone would point variables we can't directly act on. So it becomes easier to take corrective actions on the process to reach the goal. The stakes are so important that even a little improvement leads to consequent benefits, and we don't even talk about the impact on the market shares that is even more important.

7.2 Production case study

This second example is about a phenomenon that impacts the productivity of a steel plant. We have:

- a (confidential) response variable,
- $p = 145$ variables from the whole process to explain it but only $n = 100$ individuals.
- The stakes : 20% of productivity to gain on a specific product with high added value.

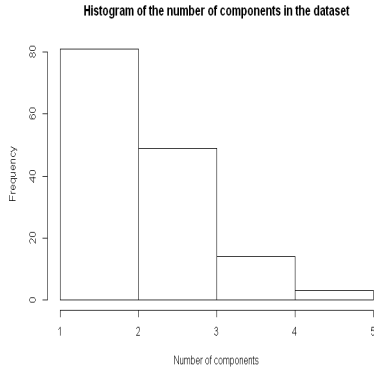


Figure 7.5: Distribution of the number of components found for each covariate.

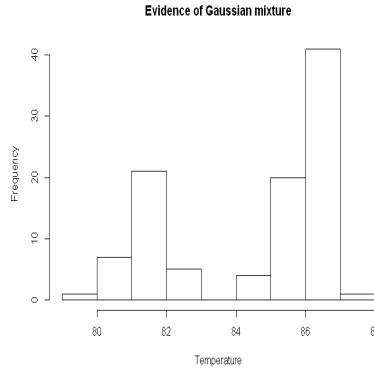


Figure 7.6: Another example of non-Gaussian real variable easily modeled by a Gaussian mixture.

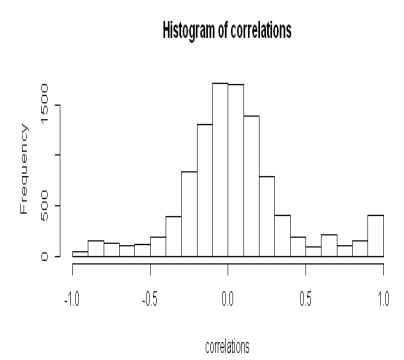


Figure 7.7: Histogram of correlations in \mathbf{X} .

CorReg found 55 sub-regressions as shown in Figure 7.8. One of them seems to be weak $R^2 = 0.17$ but is not linear (points out a link between diameter of a coil and some shape indicator).

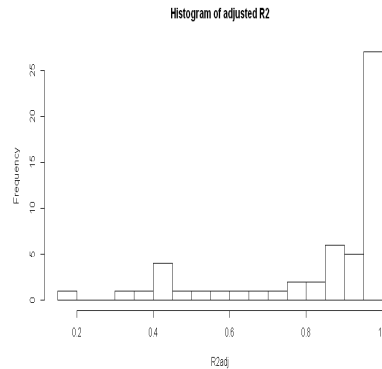


Figure 7.8: R_{adj}^2 of the 55 sub-regressions.

The response variable was binary but n was too small compared to p to use logistic regression so we have considered \mathbf{Y} as a continuous variable and then made imputation by 1 when $\hat{\mathbf{Y}} > 0.5$ and by 0 else.

In this precise case, CORREG found a structure that helped to decorrelate covariates in interpretation and to find the relevant part of the process to optimize. This product is made by a long process that requires several steel plants so it was necessary to point out the steel plant where the problem occurred.

Method	indicator	With CorReg	without CorReg
OLS	well-classified	100	56
	MSE (leave-one-out)	1.95	51 810
	complexity	91	100
LASSO	well-classified	93	93
	MSE (leave-one-out)	0.106	0.120
	complexity	27	34
Elasticnet	well-classified	84	87
	MSE (leave-one-out)	0.140	0.148
	complexity	10	13
Ridge	well-classified	88	85
	MSE (leave-one-out)	0.179	0.177
	complexity	91	146

Table 7.2: Results obtained with leave-one out cross-validation. $n = 100, p = 145$.

Part II

Further usage of the structure

Chapter 8

Missing values

Résumé: Le modèle génératif complet sur les données nous permet d'obtenir la loi des valeurs manquantes. Mais nous pouvons aller encore plus loin car la modélisation explicite des corrélations nous permet d'obtenir les lois conditionnelles de chaque valeur manquante sachant les valeurs observées. Ce chapitre présente comment nous pouvons par un simple algorithme SEM estimer les paramètres des sous-régressions dans la chaîne MCMC de recherche de structure. Enfin, nous pouvons imputer les valeurs manquantes à l'aide d'un Gibbs qui procède par imputations multiples, fournissant au passage un indicateur de précision sur l'imputation proposée.

8.1 Introduction

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. $\Theta = \{\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\}$ stands for the parameters of the Gaussian mixture followed by \mathbf{X} . $\boldsymbol{\alpha}$ 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 .

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).

We have a full generative model on \mathbf{X} with explicit dependencies within the covariates. So when a value is missing, we know its distribution but more than that, we know its conditional distribution based on observed values for the same individual. Thus we are able to make imputation and to describe the missing values with their conditional distribution. This is a positive side-effect of the explicit generative model on \mathbf{X} .

8.2 Estimation of the sub-regressions with missing values

8.2.1 The integrated likelihood

The first thing we do with \mathbf{X} is to estimate S. During the MCMC, for each candidate we have to compute the likelihood of the candidate, depending on $\boldsymbol{\alpha}$ the matrix of the sub-regression

coefficients. We start with the complete likelihood of \mathbf{X}

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}) = \prod_{i=1}^n f(\mathbf{X}_i) = \prod_{i=1}^n [f(\mathbf{X}_i^{I_r} | \mathbf{X}_i^{I_f}; \boldsymbol{\alpha}, \Theta, S) f(\mathbf{X}_i^{I_f}; \boldsymbol{\alpha}, \Theta, S)] \quad (8.1)$$

$$= \prod_{i=1}^n \left[\prod_{j \in I_r} f(x_{i,j} | \mathbf{X}_i^{I_f}; \boldsymbol{\alpha}, \Theta, S) \prod_{j \notin I_r} f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S) \right] \quad (8.2)$$

$$= \prod_{i=1}^n \left[\prod_{j \in I_r} f(x_{i,j} | \mathbf{X}_i^{I_f^j}; \boldsymbol{\alpha}, \Theta, S) \prod_{j \notin I_r} f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S) \right] \quad (8.3)$$

$$\mathcal{L}(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}) = \sum_{i=1}^n \left[\sum_{j \in I_r} \log \left(f(x_{i,j} | \mathbf{X}_i^{I_f^j}; \boldsymbol{\alpha}, \Theta, S) \right) + \sum_{j \notin I_r} \log (f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S)) \right] \quad (8.4)$$

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 (*iid* individuals, vectors of orthogonal Gaussian mixtures \mathbf{X}^{I_f} and linear combinations of these Gaussian mixtures and some Gaussian for \mathbf{X}^{I_r}) with K the number of its components.

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}_0) = \int_{\mathbf{X}_M} L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}) d\mathbf{X} = \int_{\mathbf{X}_M} \sum_{k=1}^K \pi_k \phi_k(\mathbf{X}; \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} \quad (8.5)$$

$$= \sum_{k=1}^K \pi_k \int_{\mathbf{X}_M} \phi_k(\mathbf{X}; \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} = \sum_{k=1}^K \pi_k \int_{\mathbf{X}_M} \prod_{i=1}^n \phi_k(\mathbf{X}_i; \boldsymbol{\alpha}, \Theta, S) d\mathbf{X} \quad (8.6)$$

$$= \sum_{k=1}^K \pi_k \prod_{i=1}^n \int_{\mathbf{X}_{i,M}} \phi_k(\mathbf{X}_i; \boldsymbol{\alpha}, \Theta, S) d\mathbf{X}_i = \sum_{k=1}^K \pi_k \prod_{i=1}^n \phi_k(\mathbf{X}_{i,O}; \boldsymbol{\alpha}, \Theta, S) \quad (8.7)$$

$$= \sum_{k=1}^K \pi_k \phi_k(\mathbf{X}_O; \boldsymbol{\alpha}, \Theta, S) = f(\mathbf{X}_O, \boldsymbol{\alpha}, \Theta, S) \quad (8.8)$$

To compute this likelihood, we will use the decomposition

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}_0) = f(\mathbf{X}_O; \boldsymbol{\alpha}, \Theta, S) = \prod_{i=1}^n f(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) f(\mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) \quad (8.9)$$

$$= \prod_{i=1}^n f(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) \prod_{\substack{j \in I_f \\ M_{i,j}=0}} f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S) \quad (8.10)$$

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

$$f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S) = \sum_{k=1}^{K_j} \pi_{j,k} \Phi_k(x_{i,j}; \mu_{j,k}, \Sigma_{j,k}) \quad (8.11)$$

with $K_j, \pi_{j,k}, \mu_{j,k}, \Sigma_{j,k}$ and the likelihood estimated once (for example by RMixmod [Auder et al., 2014]) before the MCMC starts.

And, $\forall(i, j)$ with $M_{i,j} = 0$ and $j \in I_r$:

$$f(x_{i,j} | \mathbf{X}_{i,O}^{I_f^j}; \boldsymbol{\alpha}, \Theta, S) = \sum_{k=1}^{K_{ij}} \pi_{ij,k} \Phi(x_{i,j}; \mu_{ij,k}, \Sigma_{ij,k}) \text{ where} \quad (8.12)$$

$$\boldsymbol{\pi}_{ij} = \bigotimes_{\substack{l \in I_f^j \\ M_{i,l}=1}} \boldsymbol{\pi}_l \text{ and } K_{ij} = |\boldsymbol{\pi}_{ij}|, \quad (8.13)$$

$$\boldsymbol{\mu}_{ij} = \sum_{\substack{l \in I_f^j \\ M_{i,l}=0}} \alpha_{l,j} x_{i,l} + \bigoplus_{\substack{l \in I_f^j \\ M_{i,l}=1}} \alpha_{l,j} \boldsymbol{\mu}_l \quad (8.14)$$

$$\Sigma_{ij} = \sigma_j^2 + \bigoplus_{\substack{l \in I_f^j \\ M_{i,l}=1}} \alpha_{i,l}^2 \Sigma_l \quad (8.15)$$

This could be easily used for imputation of the missing values in \mathbf{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 $f(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S)$ is not the product of the $f(x_{i,j} | \mathbf{X}_{i,O}^{I_f^j}; \boldsymbol{\alpha}, \Theta, S)$ if a same missing value occurs in distinct sub-regressions. Thus if every sub-regression are distinct connex component then we can use (8.12) and we have

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}_0) = \prod_{i=1}^n \prod_{\substack{j \in I_r \\ M_{i,j}=0}} f(x_{i,j} | \mathbf{X}_{i,O}^{I_f^j}; \boldsymbol{\alpha}, \Theta, S) \prod_{\substack{j \in I_f \\ M_{i,j}=0}} f(x_{i,j}; \boldsymbol{\alpha}, \Theta, S) \quad (8.16)$$

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 $f(\mathbf{X}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_{X,k}; \Sigma_{X,k})$.

$$L(\boldsymbol{\alpha}, \Theta, S; \mathbf{X}_0) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \Phi_k(\mathbf{X}_{i,O}; \boldsymbol{\alpha}, \Theta, S) \quad (8.17)$$

$$= \prod_{i=1}^n \sum_{k=1}^K \pi_k \Phi_k(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) \Phi_k(\mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) \quad (8.18)$$

$$= \prod_{i=1}^n \sum_{k=1}^K \pi_k \Phi_k(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) \prod_{\substack{j \in I_f \\ M_{i,j}=0}} \Phi_k(x_{i,j}; \mu_{j,k}, \Sigma_{j,k}) \quad (8.19)$$

Where

$$\boldsymbol{\pi} = \bigotimes_{j \in I_f} \boldsymbol{\pi}_j \text{ (Kronecker product)} \quad (8.20)$$

$$K = |\boldsymbol{\pi}| \quad (8.21)$$

$$\boldsymbol{\mu}_{X^{I_f}} = \prod_{j \in I_f} \boldsymbol{\mu}_j \text{ (Cartesian product)} \quad (8.22)$$

$$\boldsymbol{\sigma}_X = \prod_{j \in I_f} \boldsymbol{\sigma}_j \text{ (Cartesian product)} \quad (8.23)$$

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

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

$$\Phi_k(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) = \Phi_k(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k}, \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k}) \quad (8.24)$$

$$P(\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}; \boldsymbol{\alpha}, \Theta, S) = \Phi_k(\mathbf{X}_{i,O}^{I_r}; \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k}, \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k}) \quad (8.25)$$

$$\boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k} = \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r}, k} + \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r}, \mathbf{X}_{i,O}^{I_f}, k} (\boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_f}, \mathbf{X}_{i,O}^{I_f}, k})^{-1} (\mathbf{X}_{i,O}^{I_f} - \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_f}, k}) \quad (8.26)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r} | \mathbf{X}_{i,O}^{I_f}, k} = \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r}, \mathbf{X}_{i,O}^{I_r}, k} - \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_r}, \mathbf{X}_{i,O}^{I_f}, k} (\boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_f}, \mathbf{X}_{i,O}^{I_f}, k})^{-1} \boldsymbol{\Sigma}_{\mathbf{X}_{i,O}^{I_f}, \mathbf{X}_{i,O}^{I_r}, k} \quad (8.27)$$

$$\forall j \in I_r : \boldsymbol{\mu}_{\mathbf{X}_{i,O}^{I_r}} = \sum_{l \in I_f^j} \alpha_{l,j} \boldsymbol{\mu}_{l,k} \quad (8.28)$$

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

$$\text{var}_k(x_{i,j}) = \sigma_j^2 + \sum_{l \in I_f^j} \alpha_{l,j}^2 \sigma_{X^l,k}^2 \quad (8.29)$$

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

$$\text{var}_k(x_{i,j}) = \sigma_{X^j,k}^2 \quad (8.30)$$

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

$$\text{cov}_k(x_{i,j_1}, x_{i,j_2}) = \sum_{l \in I_f^{j_1} \cap I_f^{j_2}} \alpha_{l,j_1} \alpha_{l,j_2} \text{var}_k(x_{i,l}) = \sum_{l \in I_f^{j_1} \cap I_f^{j_2}} \alpha_{l,j_1} \alpha_{l,j_2} \sigma_{X^l,k}^2 \quad (8.31)$$

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

$$\text{cov}_k(x_{i,j_1}, x_{i,j_2}) = 0 \quad (8.32)$$

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

$$\text{cov}_k(x_{i,j_1}, x_{i,j_2}) = 0 \quad (8.33)$$

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

$$\text{cov}_k(x_{i,j_1}, x_{i,j_2}) = \alpha_{j_2,j_1} \sigma_{X^{j_2},k}^2 \quad (8.34)$$

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

$$\text{cov}_k(x_{i,j_1}, x_{i,j_2}) = 0 \quad (8.35)$$

8.2.2 Likelihood computation optimized

The main problem with the likelihood in its global form (8.19) is that the number of components explodes so we can't use it in practice. But in many case, it can be simplified. 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. Another strictly technical tip would be to use sparse matrix storage to avoid null value storage and useless zero multiplications. Moreover, we can look if there are missing values shared by several sub-regression. Connex component detection could be done to reduce the dimension down to strictly dependent covariates and use equation (8.12) elsewhere. We just need to compute the row-sums of the adjacency matrix G 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 (8.16). For faster computation we can

stock the vector of covariates that have missing values. So the true value of the likelihood can be computed efficiently 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 and we need to compute the likelihood for a great number of candidates. Then it is possible to use directly the simplified form of the likelihood, that can be seen as an approximation of the likelihood, not taking into account some of the dependencies but it would offer no guarantee in terms of efficiency for the MCMC.

8.2.3 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.

$$k_j = \frac{1}{n} \sum_{i=1}^n k_{i,j} \quad (8.36)$$

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

$$-2 \log P(\mathbf{X} | S) \approx BIC = -2\mathcal{L}(\mathbf{X}, S, \Theta) + |\Theta| \log(n) \quad (8.37)$$

$$= -2\mathcal{L}(\mathbf{X}, S, \Theta) + \left(\sum_{j=1}^p k_j \right) \log(n) \quad (8.38)$$

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. Or to make a compromise and penalize by $\frac{k_i \log(p)}{\log(n)}$.

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 $n - estimated$ so our proposed penalty term is an intuitive way to use a weighted penalty.

8.3 SEM

The integrated likelihood depicted above depends on α which was formerly estimated by OLS when there was no missing values. But when missing values occurs in a sub-regression we need another solution.

We use a Stochastic Expectation Maximization (SEM) algorithm [Celeux and Diebolt, 1986] to estimate α because missing values do not allow to use OLS and the log-likelihood (8.4) is not linear so a simple Expectation-Maximization (EM) would be difficult to compute.

8.3.1 Our implementation of SEM

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

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

M step: We estimate

$$\alpha^{(h+1)} = \operatorname{argmax}_{\alpha} E[\mathcal{L}(\mathbf{X}|\alpha, S, \Theta)] \quad (8.39)$$

and we can use the same method as the one for classical case without missing values (OLS, SUR, *etc.*). 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.

Another (faster but not optimal) way would be 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 SEM would then be used only once with the final structure to make imputation in \mathbf{X} before using variable selection methods like the LASSO.

8.3.2 Stochastic imputation by Gibbs sampling

We use a Gibbs sampling method to generate the missing values at the SE step. \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.

Iteration: At each iteration of the Gibbs sampler:

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

$$P(x_{i,j}|\mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}, Z; \alpha^{(h)}, \Theta, S) = P(x_{i,j}|\mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}; \alpha^{(h)}, \Theta, S) \quad (8.40)$$

$$= P(x_{i,j}|\mathbf{X}_i^{I_f^j}; \alpha^{(h)}, \Theta, S) = \mathcal{N}(\mathbf{X}_i^{I_f^j} \alpha_{I_f^j, j}^{(h)}; \sigma_j^2) \quad (8.41)$$

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

$$P(x_{i,j}|\mathbf{X}_{i,O}, \mathbf{X}_{i,\bar{M}_{i,j}}, Z; \alpha^{(h)}, \Theta, S) = P(x_{i,j}|\mathbf{X}_{i,\bar{J}}, Z_i; \alpha^{(h)}, \Theta, S) \quad (8.42)$$

$$= \mathcal{N}(\mu_{j|Z_{i,j}} + \boldsymbol{\Sigma}_{j, X_{\bar{ij}}|Z_i} \boldsymbol{\Sigma}_{X_{\bar{ij}}, X_{\bar{ij}}|Z_i}^{-1} (X_{\bar{ij}} - \boldsymbol{\mu}_{X_{\bar{ij}}|Z_i}); \sigma_{j|Z_{i,j}}^2 - \boldsymbol{\Sigma}_{j, X_{\bar{ij}}|Z_i} \boldsymbol{\Sigma}_{X_{\bar{ij}}, X_{\bar{ij}}|Z_i}^{-1} \boldsymbol{\Sigma}'_{j, X_{\bar{ij}}|Z_i}) \quad (8.43)$$

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

$$P(Z_{i,j}|\mathbf{X}, Z_{i,\bar{J}}; \Theta, \alpha, S) = P(Z_{i,j}|\mathbf{X}_i, Z_{i,\bar{J}}; \Theta, \alpha, S) = \mathcal{M}(t_{i,j,1}, \dots, t_{i,j,K_j}) \quad (8.44)$$

$$\text{where } t_{i,j,k} = \frac{\pi_{j,k} \Phi(x_{i,j}; \mu_{j,k}, \sigma_{j,k}^2)}{\sum_{l=1}^{K_j} \pi_{j,l} \Phi(x_{i,j}; \mu_{j,l}, \sigma_{j,l}^2)} \quad (8.45)$$

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. 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:

$$\begin{aligned} P(\mathbf{X}_O; \Theta, S, \alpha) &= \sum_{Z \in \mathcal{Z}} \int_{\mathbf{X}_M} \frac{P(\mathbf{X}_M, Z, \mathbf{X}_O; \Theta, \alpha, S)}{P(\mathbf{X}_M, Z | \mathbf{X}_O; \Theta, \alpha, S)} P(\mathbf{X}_M, Z | \mathbf{X}_O; \Theta, \alpha, S) dX \quad (8.46) \\ &\approx \frac{1}{Q} \sum_{q=1}^Q \frac{P(\mathbf{X}_M^{(q)}, \mathbf{X}_O, Z^{(q)}; \Theta, \alpha, S)}{P(\mathbf{X}_M^{(q)} | \mathbf{X}_O, Z^{(q)}; \Theta, \alpha, S)} \text{ by the law of large numbers} \quad (8.47) \end{aligned}$$

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

$$P(\mathbf{X}_O; \Theta, S, \alpha) \approx \frac{1}{Q} \sum_{q=1}^Q \frac{P(\mathbf{X}_M^{(q)}, \mathbf{X}_O, Z^{(q)}; \Theta, \alpha^{(q)}, S)}{P(\mathbf{X}_M^{(q)} | \mathbf{X}_O, Z^{(q)}; \Theta, \alpha^{(q)}, S)} \quad (8.48)$$

8.3.3 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 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(8.12) gives:

$$E[x_{i,j} | \alpha^{(h)}, \Theta, \mathbf{X}_O, S] = E\left[\sum_{k=1}^{k_{ij}} \pi_{ij,k} \Phi(x_{i,j} | \mu_{ij,k}, \Sigma_{ij,k}) | \alpha^{(h)}, \Theta, \mathbf{X}_O, S\right] \quad (8.49)$$

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:

$$E[x_{i,j} | \alpha^{(h)}, \Theta, \mathbf{X}_O, S] = \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\alpha^{(h)}, \Theta, \mathbf{X}_O, S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \varepsilon_k(i) - \sum_{l \in I_f^k} x_{i,l} \alpha_{l,k} \right) \right] \quad (8.50)$$

$$= \frac{1}{|r_{i,j}|} \sum_{k \in r_{i,j}} E_{|\alpha^{(h)}, \Theta, \mathbf{X}_O, S} \left[\frac{1}{\alpha_{j,k}} \left(x_{i,k} - \sum_{l \in I_f^k} x_{i,l} \alpha_{l,k} \right) \right] \quad (8.51)$$

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.

8.4 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 draw 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 more work but could be a good perspective for our method.

8.5 Numerical results on simulated datasets

8.5.1 Estimation of the sub-regression coefficients

8.5.2 Multiple imputation for the main regression

8.6 Missing values in real life

One advantage of our regression model is that it does not depend on the response variable \mathbf{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, we can keep the same \mathbf{S} and use previously computed values of $\boldsymbol{\alpha}$ as initial value for the SEM.

To be able to evaluate imputation quality we work with a real dataset but artificial missing values.

Chapter 9

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.

9.1 The model

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

$$\mathbf{Y} = \mathbf{X}^{I_r} \beta_{I_r} + \mathbf{X}^{I_f} \beta_{I_f} + \varepsilon_Y \quad (9.1)$$

$$\mathbf{X}^{I_r} = \mathbf{X}^{I_f} \alpha + \varepsilon \quad (9.2)$$

$$\mathbf{Y} = \mathbf{X}^{I_r} \underbrace{(\beta_{I_r} + \alpha \beta_{I_f})}_{\beta^*} + \varepsilon \beta_{I_r} + \varepsilon_Y \quad (9.3)$$

$$\mathbf{Y} - \mathbf{X}^{I_r} \beta^* = \varepsilon \beta_{I_r} + \varepsilon_Y \quad (9.4)$$

$$\varepsilon = \mathbf{X}^{I_r} - \mathbf{X}^{I_f} \alpha \quad (9.5)$$

So we introduce a plug-in model

$$\underbrace{\mathbf{Y} - \mathbf{X}^{I_r} \hat{\beta}^*}_{\tilde{\mathbf{Y}}} = \underbrace{(\mathbf{X}^{I_r} - \mathbf{X}^{I_f} \hat{\alpha})}_{\tilde{\mathbf{X}}} \beta_{I_r} + \varepsilon_Y \quad (9.6)$$

$$(9.7)$$

That allows us to estimate β_{I_r} with a classical linear model based on previous estimations of β^* and α . Then we have a model with a smaller noise

$$\mathbf{Y} = \mathbf{X}^{I_r} \beta^* + \varepsilon \hat{\beta}_{I_r} + \varepsilon_Y \quad (9.8)$$

9.2 Interpretation and latent variables

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

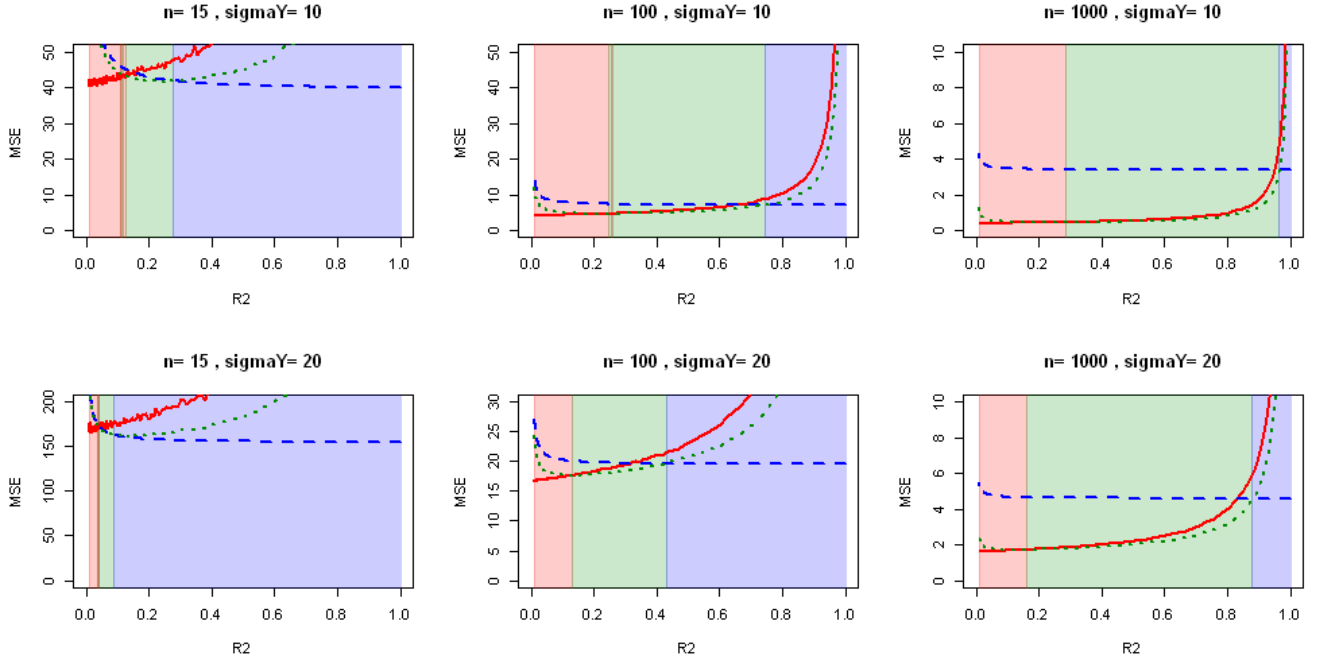


Figure 9.1: MSE of OLS (plain red) and CorReg marginal(blue dashed) and CorReg full (green dotted) estimators for varying R^2 of the sub-regression, n and σ_Y .

9.3 Consistency

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 $\mathbf{X}^{I_f}, \mathbf{X}^{I_r}, \epsilon_Y, \epsilon_X i.i.d. \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and then

$$\mathbf{X}_3 = \frac{2}{3}\mathbf{X}_1 + \frac{2}{3}\mathbf{X}_2 + \frac{1}{3}\epsilon_X \text{ and}$$

$$\mathbf{Y} = 2\mathbf{X}_1 + 3\mathbf{X}_2 + \epsilon_Y.$$

We compare consistencies of complete, marginal and full plug-in 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 $\hat{\alpha}$ is used) and with the structure found by the Markov chain after a few seconds.

True S was found 340 times on 1000 tries (model is not identifiable because \mathbf{X}^j are all Gaussian).

	Classical LASSO	CorReg + marginal LASSO	CorReg + full plug-in LASSO
True S	1.006479	1.005468	1.006093
\hat{Z}	1.006479	1.884175	1.006517

Table 9.1: MSE observed on a validation sample (1000 individuals)

We observe as we hoped that our marginal model is better when using true S (coercing real zeros) and that marginal with \hat{S} is penalized (coercing wrong coefficients to be zeros when true S is not found). But the main point is that the plug-in model stay better than the classical one with the true S and corrects enough the marginal model to follow the classical LASSO closely when using \hat{S} . And when we look at the consistency :

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).

	Classical LASSO	CorReg + marginal LASSO	CorReg + full plug-in LASSO
True S	0	1000	830
\hat{S}	0	340	621

Table 9.2: Number of consistent model found (\mathbf{Y} depending on $\mathbf{X}_1, \mathbf{X}_2$ and only them) on 1000 tries

We also made the same experiment but with $\mathbf{X}_1, \mathbf{X}_2$ (and consequently \mathbf{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 Gaussian mixture models are easier to identify.

	Classical LASSO	CorReg + marginal LASSO	CorReg + full plug-in LASSO
True S	1.571029	1.569559	1.570801
\hat{S}	1.005402	1.465768	1.005066

Table 9.3: MSE observed on a validation sample (1000 individuals)

And when we look at the consistency :

	Classical LASSO	CorReg + marginal LASSO	CorReg + full plug-in LASSO
True S	0	1000	789
\hat{S}	0	714	608

Table 9.4: 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).

9.4 Numerical results

Chapter 10

CorReg: the concept

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). The objective of CorReg is to bring recent statistical tools to engineers. Thus it will be made available in Microsoft Excel for the end of the year 2014, probably using Basic Excel R Toolkit(BERT¹). It also provides some small scripts in functions to obtain graphical representations and basic statistics with legends for non-statistician with only one command line (or macro button in Excel).

¹<https://github.com/StructuredDataLLC/Basic-Excel-R-Toolkit>

Chapter 11

Conclusion and perspectives

11.1 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 look 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).

11.2 Perspective

11.2.1 Non-linear regression

Polynomial regression, logistic regression [Hosmer and Lemeshow, 2000], *etc.* could be improved by a method like this.

11.2.2 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.

11.2.3 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).

11.2.4 Missing values in classical methods

The full generative approach could be used to manage missing values without imputation for many classical methods. It can notably be used for clustering and not only in response variable

prediction context.

11.2.5 Interpretation improvements

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

Bibliography

- [Abdi, 2003] Abdi, H. (2003). Partial least squares regression (pls-regression).
- [Amemiya, 1985] Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.
- [Anderson, 1984] Anderson, T. W. (1984). Estimating linear statistical relationships. *The Annals of Statistics*, pages 1–45.
- [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.
- [Arlot et al., 2010] Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.
- [Auder et al., 2014] Auder, B., Lebre, R., Iovleff, S., and Langrognet, F. (2014). *Rmixmod: An interface for MIXMOD*. R package version 2.0.2.
- [Biernacki et al., 2006] Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600.
- [Biggs, 1993] Biggs, N. (1993). *Algebraic graph theory*. Cambridge University Press.
- [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.
- [Bondy and Murty, 1976] Bondy, J. and Murty, U. (1976). *Graph theory with applications*, volume 290. Macmillan London.
- [Breiman, 1984] Breiman, L. (1984). *Classification and regression trees*. Chapman & Hall/CRC.
- [Breiman and Friedman, 1997] Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54.
- [Brito and Pearl, 2006] Brito, C. and Pearl, J. (2006). Graphical condition for identification in recursive sem.
- [Cameron and Trivedi, 2005] Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.
- [Casella and George, 1992] Casella, G. and George, E. (1992). Explaining the gibbs sampler. *American Statistician*, pages 167–174.

- [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.
- [Chauveau, 1995] Chauveau, D. (1995). A stochastic {EM} algorithm for mixtures with censored data. *Journal of Statistical Planning and Inference*, 46(1):1 – 25.
- [Chib and Greenberg, 1995] Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *American Statistician*, pages 327–335.
- [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*.
- [Cule, 2014] Cule, E. (2014). *ridge: Ridge Regression with automatic selection of the penalty parameter*. R package version 2.1-3.
- [Cule and De Iorio, 2013] Cule, E. and De Iorio, M. (2013). Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genetic epidemiology*, 37(7):704–714.
- [d’Aspremont et al., 2008] d’Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66.
- [Davidson and MacKinnon, 1993] Davidson, R. and MacKinnon, J. (1993). Estimation and inference in econometrics. *Oxford University Press Catalogue*.
- [Dempster, 1972] Dempster, A. (1972). Covariance selection. *Biometrics*, pages 157–175.
- [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.
- [Diebolt and Ip, 1996] Diebolt, J. and Ip, E. (1996). A stochastic em algorithm for approximating the maximum likelihood estimate. *Markov chain Monte Carlo in practice*.
- [Dodge and Rousson, 2004] Dodge, Y. and Rousson, V. (2004). Analyse de régression appliquée: manuel et exercices corrigés (coll. eco sup,). *Recherche*, 67:02.
- [Edwards, 1984] Edwards, A. W. (1984). *Likelihood T*. CUP Archive.
- [Efron, 1979] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26.
- [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.
- [Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, volume 57. CRC press.

- [Er et al., 2013] Er, M. J., Shao, Z., and Wang, N. (2013). A systematic method to guide the choice of ridge parameter in ridge extreme learning machine. In *Control and Automation (ICCA), 2013 10th IEEE International Conference on*, pages 852–857. IEEE.
- [Eubank, 1999] Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.
- [Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- [Fausett, 1994] Fausett, L. (1994). Fundamentals of neural networks: architectures, algorithms, and applications. *Englewood Cliffs*.
- [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.
- [Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical report, Stanford University.
- [Friedman et al., 2000] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620.
- [Gareth M.J. and R., 2013] Gareth M.J., C. P. and R., P. (2013). The constrained lasso. URL: <http://www-bcf.usc.edu/~gareth/research/CLassoFinal.pdf>.
- [Geladi and Kowalski, 1986] Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17.
- [George, 2000] George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308.
- [George and McCulloch, 1993] George, E. and McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, pages 881–889.
- [Gilks et al., 1996] Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*, volume 2. CRC press.
- [Goldberger, 1984] Goldberger, A. S. (1984). Reverse regression and salary discrimination. *Journal of Human Resources*, pages 293–318.
- [Gower and Ross, 1969] Gower, J. and Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(1):54–64.
- [Graham and Hell, 1985] Graham, R. and Hell, P. (1985). On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7(1):43–57.
- [Greene, 1984] Greene, W. H. (1984). Reverse regression: The algebra of discrimination. *Journal of Business and Economic Statistics*, 2(2):pp. 117–120.

- [Grinstead and Snell, 1997] Grinstead, C. M. and Snell, J. (1997). *Introduction to probability*. American Mathematical Society.
- [Härdle, 1990] Härdle, W. (1990). *Applied nonparametric regression*, volume 27. Cambridge Univ Press.
- [Heckerman et al., 1995] Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243.
- [Hoerl and Kennard, 1970] Hoerl, A. and Kennard, R. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, pages 69–82.
- [Hosmer and Lemeshow, 2000] Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression*, volume 354. Wiley-Interscience.
- [Hox, 1998] Hox, J. (1998). Multilevel modeling: When and why. In *Classification, data analysis, and data highways*, pages 147–154. Springer.
- [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.
- [Isobe et al., 1990] Isobe, T., Feigelson, E., Akritas, M., and Babu, G. (1990). Linear regression in astronomy. *The astrophysical journal*, 364:104–113.
- [Jackson, 2005] Jackson, J. E. (2005). *A user’s guide to principal components*, volume 587. John Wiley & Sons.
- [Jensen and Nielsen, 2007] Jensen, F. and Nielsen, T. (2007). *Bayesian networks and decision graphs*. Springer Verlag.
- [Karalic, 1992] Karalic, A. (1992). Linear regression in regression tree leaves. In *In Proceedings of ECAI-92*, pages 440–441. John Wiley and Sons.
- [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.
- [Kiebel and Holmes, 2003] Kiebel, S. and Holmes, A. (2003). The general linear model. *Human brain function*, 2:725–760.
- [Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145.
- [Le Cessie and Van Houwelingen, 1992] Le Cessie, S. and Van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied statistics*, pages 191–201.
- [Leamer, 1978] Leamer, E. E. (1978). Least-squares versus instrumental variables estimation in a simple errors in variables model. *Econometrica: Journal of the Econometric Society*, pages 961–968.
- [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.
- [Li, 1991] Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, pages 316–327.

- [Little, 1992] Little, R. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.
- [Longford, 2012] Longford, N. (2012). A revision of school effectiveness analysis. *Journal of Educational and Behavioral Statistics*, 37(1):157–179.
- [Maas and Hox, 2004] Maas, C. J. and Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2):127–137.
- [Marquardt and Snee, 1975] Marquardt, D. and Snee, R. (1975). Ridge regression in practice. *American Statistician*, pages 3–20.
- [Massart and Picard, 2007] Massart, P. and Picard, J. (2007). *Concentration inequalities and model selection*, volume 1896. Springer.
- [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.
- [McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models.
- [McLachlan and Krishnan, 2007] McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.
- [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.
- [Moerbeek et al., 2003] Moerbeek, M., van Breukelen, G. J., and Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of clinical epidemiology*, 56(4):341–350.
- [Montgomery et al., 2012] Montgomery, D., Peck, E., and Vining, G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.
- [Moret and Shapiro, 1991] Moret, B. and Shapiro, H. (1991). An empirical analysis of algorithms for constructing a minimum spanning tree. *Algorithms and Data Structures*, pages 400–411.
- [Nelder and Baker, 1972] Nelder, J. A. and Baker, R. (1972). *Generalized linear models*. Wiley Online Library.
- [Pearl, 1998] Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.
- [Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning, and inference*, volume 47. Cambridge Univ Press.
- [Penrose, 1955] Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51:406–413.
- [Pierro and Wei, 2000] Pierro, A. R. D. and Wei, M. (2000). Some new properties of the equality constrained and weighted least squares problem. *Linear Algebra and its Applications*, 320:145 – 165.

- [Quinlan, 1986] Quinlan, J. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.
- [R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [Radchenko and James, 2008] Radchenko, P. and James, G. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, 103(483).
- [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.
- [Reiersol, 1950] Reiersol, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica*, 18(S 375).
- [Roberts and Rosenthal, 2001] Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367.
- [Saporta, 2006] Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.
- [Saracco et al., 1999] Saracco, J., Larramendy, I., and Aragon, Y. (1999). La regression inverse par tranches ou méthode sir: presentation générale. *La revue de Modulad*, (22):21–39.
- [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.
- [Wickens, 2004] Wickens, T. (2004). The general linear model. *Institute for Pure and Applied Mathematics*. URL: http://www.ipam.ucla.edu/publications/mbi2004/mbi2004_5017.pdf.
- [Witten et al., 2011] Witten, D., Friedman, J., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.
- [Witten and Tibshirani, 2009] Witten, D. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636.
- [Xu et al., 2013] Xu, H., Eis, D., and Ramadge, P. (2013). The generalized lasso is reducible to a subspace constrained lasso. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3268–3272. IEEE.
- [Yengo et al., 2012] Yengo, L., Jacques, J., Biernacki, C., et al. (2012). Variable clustering in high dimensional linear regression models.

- [Yuan et al., 2007] Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.
- [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.
- [Zhou and Lange, 2013] Zhou, H. and Lange, K. (2013). A path algorithm for constrained estimation. *Journal of Computational and Graphical Statistics*, 22(2):261–283.
- [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.