CorReg: Decorrelating variables for linear regression

Clément THERY

March 20, 2014

To my sons,

Contents

Chapter 1

Abstracts

Chapter 2

Acknowledgments

Chapter 3

The industrial context

Chapter 4

State of the art

4.1 Ordinary least squares and associated problems

4.2 Dimension reduction

4.2.1 OSCAR: Octogonal Shrinkage and Clustering Algorithm for Regression

4.2.2 CLERE: CLusterwise Effect REgression

4.2.3 Spike and Slab

4.3 Penalized models

4.3.1 Ridge regression

4.3.2 LASSO: Least Absolute Shrinkage and Selection Operator

4.3.3 Adaptative LASSO and Random LASSO

4.3.4 Elasticnet

4.4 Multiple Equations

4.4.1 SEM and Path Analysis

4.4.2 SUR: Seemingly Unrelated Regression

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

Part I

CorReg : the concept

Chapter 5

Decorrelating covariates by a generative model

5.1 Generative model

5.2 Properties

5.2.1 general properties

5.2.2 Identifiability

5.3 About grouping effect

Chapter 6

Estimation of the Structure of subregression by MCMC

6.1 How to compare structures ?

6.1.1 Bayesian criterion for quality

6.1.2 Some indicators for proximity

6.2 Neighbourhood

6.2.1 Classical

6.2.2 Active relaxation of the constraints

6.3 The walk

6.4 Numerical results

Part II

Further usage of the structure

Chapter 7

Taking back the residuals

7.1 The model

7.2 Properties

7.3 Consistency

7.3.1 Consistency Issues

Consistency issues of the LASSO are well known and Zhao [4] gives a very simple example to illustrate it. We have taken the same example to show how our method is more consistent. Here $p = 3$ and $n = 1000$.We define $X_1, X_2, \varepsilon_Y, \varepsilon_X i.i.d. \sim \mathcal{N}(0,1)$ and then $X_3 = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}\varepsilon_X$ and $Y = 2X_1 + 3X_2 + \varepsilon_Y$. We compare consistencies of complete,explicative and predictive model with LASSO (and LAR) for selection. It happens that the algorithm don't find the true structure but a permuted one so we also look at the results obtained with the true S (but $\hat{B}$ is used) and with the structure found by the Markov chain after a few seconds.

True S is found 340 times on 1000 tries.

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

Table 7.1: MSE observer on a validation sample (1000 individuals)

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

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

Table 7.2: number of consistent model found (Y depending on X_1, X_2 and only them) on 1000 tries

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

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

And when we look at the consistency :

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

Table 7.3: MSE observed on a validation sample (1000 individuals)

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

Table 7.4: 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).

7.4 Numerical results

Chapter 8

Missing values

Missing values are a very recurrent issue in industry. We note M the binary matrix indicating whereas a value is missing (1) or not (0).

8.1 How to manage missing values in the MCMC ?

8.1.1 Position of the missing value

In the MCMC we need to compute the likelihood of the structure. When missing values occurs, we restrict the likelihood to the known values. So we look at each non-missing value separately. Each value is seen as different random variable noted $x_{i,j}$. But the structure itself makes things more complicated because known values are not all *iid* (if $j \in I_2$). For a given structure S, missing values can imply different consequences according to their position in the dataset. To compute the likelihood of a value $x_{i,j}$ in the dataset :

- if $M_{i,j} = 1$: $x_{i,j}$ is not considered because we restrict the likelihood to known values.
- else if $j \in I_1$: like in previous method, we use the density estimated (*e.g.* a Gaussian Mixture model estimated by MIXMOD) for X^j. Values in X^j are *iid*.
- else :

$$x_{i,j}|X_{i,I_1^j} \sim \sum_{\substack{1\leq k\leq p \\ B_k^j\neq 0}} x_{i,k}B_k^j + \mathcal{N}(0;\sigma_j) \tag{8.1}$$

$$= \sum_{\substack{1\leq k\leq p \\ B_k^j\neq 0 \\ M_{i,k}\neq 0}} x_{i,k}B_k^j + \mathcal{N}\Big(\sum_{\substack{1\leq k\leq p \\ B_k^j\neq 0 \\ M_{i,k}= 0}} x_{i,k}B_k^j \ ; \ \sigma_j\Big) \tag{8.2}$$

$$\tag{8.3}$$

8.1.2 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 $j \in I_2$, the number of parameters is not the same for each individual depending whether or not $M_{i,j} = 0$. But the penalty (for $\psi = BIC$) can't be added at the individual level (because $\log(1) = 0$ so it would be annihilated).

To penalize models that suppose dependencies based only on a few individuals, we propose to use the mean of the complexities obtained for a given covariate. Thus if a structure is only touched by one missing value the penalty will be smaller than another same shaped structure but with more missing values implied. Another way would be to use $\psi = RIC$ (see [2]) so the complexity is associated with

$\log(p)$ and can be added individually. Another idea would be to make a compromise and penalize by $\frac{k_i \log(p)}{\log(n)}$.

8.1.3 Parameters estimation

Estimating B with missing values is just estimating independent regressions with missing values. We have seen in equation (8.2) that we know the expression of this density for a given B. So it's just about maximizing the likelihood of this density on B. This can be done with an Expectation-Maximization (EM) [1] or one of its extensions [3]. But estimation of B is the mast critical part of the MCMC in terms of computational time so it could be dangerous to put there another iterative algorithm. Alternatives does exist :

- Because sub-regression are supposed to be simple, we could imagine to use estimate each column of B with full submatrices of X. When relying on too much missing values, $\hat{B}$ would be a bad candidates and then penalized directly by the likelihood (and it could be a good thing). Computational cost would be reduced significantly.
- To estimate B (and not for the global likelihood) we could use data imputation (by the mean) and then obtain a full matrix but still ignoring missing values when estimating the likelihood.

8.2 Missing values in the main regression

8.2.1 explicative

The reduced model (explicative one) is just a linear regression without structure so the method used for $\hat{B}$ can also be used here.

8.2.2 predictive

Chapter 9

CorReg: the package and its application in steel industry

9.1 CorReg package for R

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

9.2 Application in steel industry

9.2.1 The dataset

9.2.2 Found Structure

9.2.3 Results

Chapter 10

Conclusion and perspectives

Chapter 11

References

Bibliography

[1] Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[2] Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.

[3] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[4] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

Chapter 12

Appendices

12.1 Graphs and CorReg

12.1.1 Matricial notations

12.1.2 Properties

12.2 Mixture models

12.2.1 Linear combination

12.2.2 Industrial examples