CorReg: Linear regression with highly correlated covariates

Clément THERY

May 19, 2014

To my sons,

Contents

Chapter 1

Abstracts

Chapter 2

Acknowledgments

Chapter 3

The industrial context

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

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

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

Chapter 4

State of the art

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

4.1 Ordinary least squares and associated problems

Linear regression is defined by this simple equation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y \tag{4.1}$$

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

4.2 Penalized models

4.2.1 Ridge regression

Ridge regression [7] proposes a biased estimator that can be written in terms of a parametric L_2 penalty:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_2^2 \leq \lambda \text{ with } \lambda > 0 \tag{4.2}$$

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

4.2.2 LASSO: Least Absolute Shrinkage and Selection Operator

[13] [12] [4] [18][17]The Least Absolute Shrinkage and Selection Operator (LASSO [12]) consists in a shrinkage of the regression coefficients based on a λ parametric L_1 penalty to obtain zeros in $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \text{argmin} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right\} \text{ subject to } \| \boldsymbol{\beta} \|_1 \leq \lambda \text{ with } \lambda > 0 \tag{4.3}$$

The Least Angle Regression (LAR [4]) 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.

4.2.3 Adaptative LASSO and Random LASSO

[19][14]Some recent variants of the LASSO do exist for the choice of the penalization coefficient like the adaptative LASSO [19] or the random LASSO [14]. But LASSO also faces consistency problems [18] when confronted with correlated covariates.

4.2.4 Elasticnet

[20] Elastic net [20] is a method developed to be a compromise between Ridge regression and the LASSO:

$$\hat{\boldsymbol{\beta}} = (1+\lambda_2)\,\text{argmin}\left\{\parallel \boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\parallel_2^2\right\},\ \text{subject to } (1-\alpha)\parallel\boldsymbol{\beta}\parallel_1+\alpha\parallel\boldsymbol{\beta}\parallel_2^2\leq t \text{ for some } t \qquad (4.4)$$

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

4.2.5 OSCAR: Octogonal Shrinkage and Clustering Algorithm for Regression

Like elasticnet, OSCAR [1] uses combination of two norms for its penalty. Here the objective is to group covariates with the same effect (by a pairwise L_∞ norm) and give them exactly the same coefficient (reducing the dimension) with a simultaneous variable selection (implied by the L_1 norm).

$$\hat{\boldsymbol{\beta}} = \text{argmin}_\beta \parallel \boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\parallel_2^2 \ \text{ subject to } \sum_{j=1}^{p}|\beta_j| + c\sum_{j<k}\max(|\beta_j|,|\beta_k|)\leq\lambda \qquad (4.5)$$

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

4.3 Modeling the parameters

4.3.1 CLERE: CLusterwise Effect REgression

[15]The CLusterwise Effect REgression (CLERE [15]) describes the β_j no longer as fixed effect parameters but as unobserved independant random variables with grouped β_j following a Gaussian Mixture distribution. The idea is to hope that the model have a small number of groups of covariates and that the mixture will have few enough components to have a number of parameters to estimate significantly lower than p. In such a case, it improves interpretability and ability to yeld reliable prediction with a smaller variance on $\hat{\boldsymbol{\beta}}$.

4.3.2 Spike and Slab

[6]Spike and Slab variable selection [6] 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.

4.4 Multiple Equations

4.4.1 SEM and Path Analysis

4.4.2 SUR: Seemingly Unrelated Regression

[16]

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

[2]

4.4.4 Selvarclust: Linear regression within covariates for clustering

[8] 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 [11] even it is known to be often unstable [10]

We provide an specific MCMC algorithm with the ability to have redundant covariates in the true 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 [18] gives a very simple example to illustrate it. We have taken the same example to show how our method is more consistent. Here $p = 3$ and $n = 1000$.We define $\boldsymbol{X}_f, \boldsymbol{X}_r, \boldsymbol{\varepsilon}_Y, \boldsymbol{\varepsilon}_X i.i.d. \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and then $X_3 = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}\varepsilon_X$ and $Y = 2X_1 + 3X_2 + \varepsilon_Y$. We compare consistencies of complete,explicative and predictive model with LASSO (and LAR) for selection. It happens that the algorithm don't find the true structure but a permuted one so we also look at the results obtained with the true S (but $\hat{B}$ is used) and with the structure found by the Markov chain after a few seconds.

True S is found 340 times on 1000 tries.

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

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

$$g(\boldsymbol{X}|\Theta) = \int_{\boldsymbol{X}_M} f(\boldsymbol{X}|\Theta) d\boldsymbol{X} \tag{8.1}$$

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 $x_{i,j}$ for the i^{th} value of $\boldsymbol{X}^j$ the observed values. But the structure itself makes things more complicated because known values are not all *iid* (if $j \in I_r$). For a given structure S, missing values can imply different consequences according to their position in the dataset. We decompose the joint distribution $g(\boldsymbol{X}|\Theta) = g(\boldsymbol{X}_r|\Theta, \boldsymbol{X}_f)g(\boldsymbol{X}_f|\Theta)$. To compute the likelihood of a value $x_{i,j}$ in the dataset, if $\boldsymbol{M}_{i,j} = 1$: $x_{i,j}$ is not considered because we restrict the likelihood to known values (integrated likelihood). else if $j \in I_f$: like in previous method, we use the density estimated (*e.g.* a Gaussian Mixture model estimated by MIXMOD) for $\boldsymbol{X}^j$. Values of $\boldsymbol{X}^j$ are *iid*:

$$g(\boldsymbol{x}_{i,j}|\Theta) = f(\boldsymbol{x}_{i,j}|\Theta) = \sum_{k=1}^{k_j} \pi_{j,k} \Phi(\boldsymbol{x}_{i,j}|\mu_{j,k}, \Sigma_{j,k}) \tag{8.2}$$

with Θ estimated by Mixmod and else, $\forall j \in I_r$:

$$g(x_{i,j}|\boldsymbol{X}_{i,f}, \Theta) = g(x_{i,j}|\boldsymbol{X}_{i,I_f^j}, \Theta) = \sum_{k=1}^{k_{ij}} \pi_{ij,k} \Phi(x_{i,j}|\mu_{ij,k}, \Sigma_{ij,k}) \text{ where} \tag{8.3}$$

$$\boldsymbol{\pi}_{ij} = \bigotimes_{\substack{l \in I_f^j \\ \boldsymbol{M}_{i,l}=1}} \boldsymbol{\pi}_l \text{ and } k_{ij} = |\pi_{ij}|, \tag{8.4}$$

$$\boldsymbol{\mu}_{ij} = \sum_{\substack{l \in I_f^j \\ \boldsymbol{M}_{i,l}=0}} \alpha_{l,j} x_{i,l} + \bigoplus_{\substack{l \in I_f^j \\ \boldsymbol{M}_{i,l}=1}} \alpha_{l,j} \boldsymbol{\mu}_l \tag{8.5}$$

$$\boldsymbol{\Sigma}_{ij} = \sigma_j^2 + \bigoplus_{\substack{l \in I_f^j \\ \boldsymbol{M}_{i,l}=1}} \alpha_{i,l}^2 \boldsymbol{\Sigma}_l \tag{8.6}$$

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

To penalize models that suppose dependencies based only on a few individuals, we propose to use the mean of the complexities obtained for a given covariate. Thus if a structure is only touched by one missing value the penalty will be smaller than another same shaped structure but with more missing values implied. Another way would be to use $\psi = RIC$ (see [5]) 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 Estimation of the coefficients in each regression

Estimating the $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}$ with missing values is just estimating independent regressions with missing values. We have seen in equation (8.3) that we know the expression of this density for a given the $\boldsymbol{\alpha}_j$. So it's just about maximizing the likelihood of this density on the $\boldsymbol{\alpha}_j$. This can be done with an Expectation-Maximization (EM) algorithm [3] or one of its extensions [9]. But estimation of the $\boldsymbol{\alpha}_j$ is the most critical part of the MCMC in terms of computational time so it could be a bad idea to put there another iterative algorithm. Alternatives does exist :

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

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

$$\boldsymbol{X}^{(h)} = E(\boldsymbol{X}|\boldsymbol{X}_{\bar{M}}, \Theta^{(h)}) \tag{8.7}$$

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

$$\boldsymbol{X}^{(h)_{i,j}} = E(\boldsymbol{X}_{i,j}|\boldsymbol{X}_{\bar{M}}, \Theta^{(h)}) = \sum_{k=1}^{k_{ij}} \pi_{ij,k}\mu_{ij,k}^{(h)} \tag{8.8}$$

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

$$E(\boldsymbol{X}|\Theta) = \boldsymbol{X}^{(h)} \text{ done by OLS} \tag{8.9}$$

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

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 estimating $\boldsymbol{\beta}$ is like estimating the $\boldsymbol{\alpha}_j$ and the same methods can be used. An EM algorithm would be prefered because

this estimation is out of the MCMC, will be computed only one time and is the final objective where we want to minimize the error. The E step is the same as during the MCMC because we can take benefits from the structure. Where others will just fill $\boldsymbol{X}^j$ by their means, we can use when $j \in I_r$ the conditionnal mean $E(X^j|X_f, \boldsymbol{\alpha}^j)$ defined in 8.8.

8.2.2 predictive

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

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

[2] Mary-Huard T. Chiquet J. and Robin S. Multi-trait genomic selection via multivariate regression with structured regularization. *Proceedings of MLCB/NIPS'13 workshop*, 2013.

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

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

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

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

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

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

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

[10] Alan Miller. *Subset selection in regression*. CRC Press, 2002.

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

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

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

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

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

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

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

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

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

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

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