CorReg

Clément Théry, Christophe Biernacki, Gaétan Loridant

ArcelorMittal Dunkerque, Université de Lille 1,équipe MΘdal Inria

February 8, 2015

1. Steel industry databases.
2. Goal: To understand and prevent quality problems on finished product, knowing the whole process, without a priori.






Regression

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_Y^2 \boldsymbol{I}_n)$

OLS

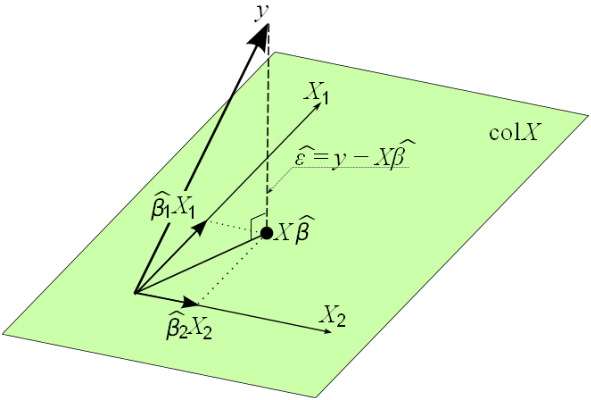

Figure: Multiple linear regression with Ordinary Least Squares seen as a projection on the $d-$dimensional hyperplane spanned by the regressors $\boldsymbol{X}$. Public domain image.

OLS

β can be estimated by $\hat{\beta}$ with Ordinary Least Squares (OLS), that is the unbiased maximum likelihood estimator [Saporta, 2006, Dodge and Rousson, 2004]:

$$\hat{\beta}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{2}$$

with variance matrix

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma_Y^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

In fact it is the Best Linear Unbiased Estimator (BLUE). The theoretical MSE is given by

$$\mathrm{MSE}(\hat{\beta}_{OLS}) = \sigma_Y^2 \,\mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Running example

$\boldsymbol{X}^1, \boldsymbol{X}^2, \boldsymbol{X}^4, \boldsymbol{X}^5 \sim \mathcal{N}(0,1)$ and $\boldsymbol{X}^3 = \boldsymbol{X}^1 + \boldsymbol{X}^2 + \varepsilon_1$ where $\varepsilon_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2 \boldsymbol{I}_n)$.

Two *scenarii* for $\boldsymbol{Y}$:

$\beta = (1,1,1,1,1)'$ and $\sigma_Y \in \{10, 20\}$.

It is clear that $\boldsymbol{X}'\boldsymbol{X}$ will become more ill-conditioned as σ_1 gets smaller. R^2 stands for the coefficient of determination which is here:

$$R^2 = 1 - \frac{\mathrm{Var}(\varepsilon_1)}{\mathrm{Var}(\boldsymbol{X}^3)} \tag{4}$$

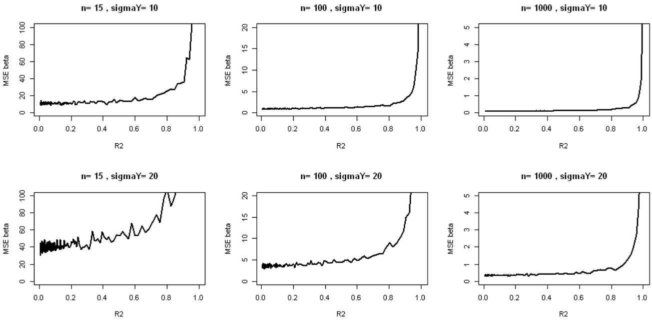

Figure: Evolution of observed Mean Squared error on $\hat{\beta}_{OLS}$ with the strength of the correlations for various sample sizes and strength of regression. $d = 5$ covariates (running example).

Ridge Regression

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

$$\hat{\beta} = \mathrm{argmin}_{\beta} \left\{ \| \boldsymbol{Y} - \boldsymbol{X}\beta \|_2^2 \right\} \text{ subject to } \| \beta \|_2^2 \leq \eta \text{ with } \eta > 0 \tag{5}$$

But this penalty is not guided by correlations. The solution of the ridge regression is given by

$$\hat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X} - \lambda \boldsymbol{I}_n\right)^{-1} \boldsymbol{X}'\boldsymbol{Y} \tag{6}$$

Methods do 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].

Ridge Regression

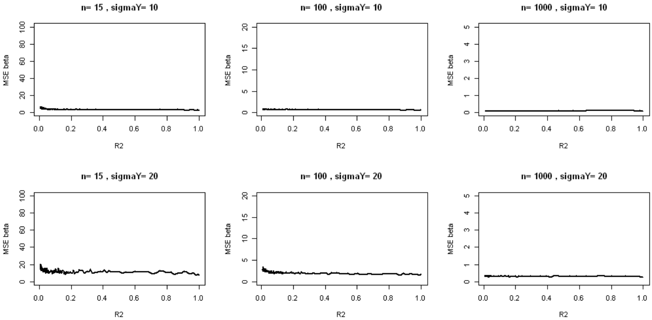

Figure: Evolution of observed Mean Squared error on $\hat{\beta}_{ridge}$ with the strength of the correlations for various sample sizes and strength of regression. $d = 5$ covariates.

LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO, [Tibshirani, 1996] and [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:

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

LASSO

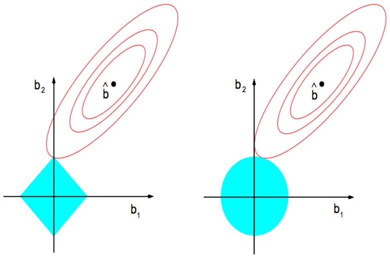

Figure: Geometric view of the Penalty for the LASSO (left) compared to ridge regression (right) as shown in the book from Hastie [Hastie et al., 2009]

Figure shows the contour of error (red) and constraint function (blue). The axis stands for the regression coefficients.

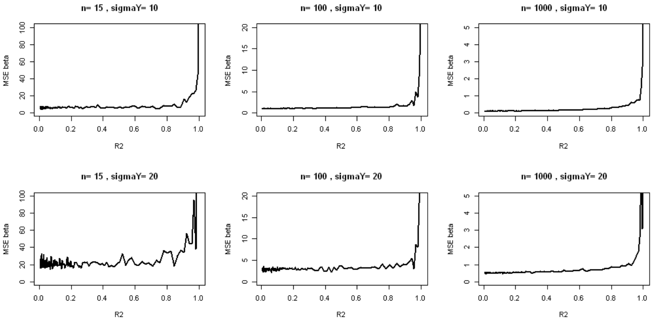

Figure: Evolution of observed Mean Squared error on $\hat{\beta}_{lar}$ with the strength of the correlations for various sample sizes and strength of regression. $d = 5$ covariates.

`lars` package on CRAN ([Hastie and Efron, 2013]).

SEM

Modélisation de la structure mais à la main et aucun impact sur l'estimation

Selvarclust

Semble très bien mais n'aboutit pas vers la régression donc on le prolonge en CorReg

Hypothesis 1

There are $d_r \geq 0$ "sub-regressions", each sub-regression $j = 1, \ldots, d_r$ having the covariate $\boldsymbol{X}^{J_r^j}$ as *response* variable ($J_r^j \in \{1, \ldots, p\}$ and $J_r^j \neq J_r^{j'}$ if $j \neq j'$) and having the $d_p^j > 0$ covariates $\boldsymbol{X}^{J_p^j}$ as *predictor* variables ($J_p^j \subset \{1, \ldots, d\} \backslash J_r^j$ and $d_p^j = |J_p^j|$ the cardinal of J_p^j):

$$\boldsymbol{X}^{J_r^j} = \boldsymbol{X}^{J_p^j} \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}_j, \tag{7}$$

where $\boldsymbol{\alpha}_j \in \mathbb{R}^{d_r^j}$ ($\alpha_j^h \neq 0$ for all $j = 1, \ldots, d_r$ and $h = 1, \ldots, d_p^j$) and $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2 \boldsymbol{I})$.

Hypothesis 2

the response covariates and the predictor covariates are totally disjoint: for any sub-regression $j = 1, \dots, d_r$, $J_p^j \subset J_f$ where $J_r = \{J_r^1, \dots, J_r^{d_r}\}$ is set of all response covariates and $J_f = \{1, \dots, d\} \setminus J_r$ is the set of all *non* response covariates of cardinal $d_f = d - d_r = |J_f|$. We call this hypothesis the uncrossing rule. Then:

$$\boldsymbol{Y} = \boldsymbol{X}_f \beta_f + \boldsymbol{X}_r \beta_r + \varepsilon_Y. \tag{8}$$

Hypotheses 3

We assume that all errors ε_Y and ε_j ($j = 1, \dots, d_r$) are *mutually independent*. It implies in particular that conditional response covariates $\{\boldsymbol{X}^{J_r^j} | \boldsymbol{X}^{J_p^j}, \boldsymbol{S}; \boldsymbol{\alpha}_j, \sigma_j^2\}$ are *mutually independent*:

$$\mathbb{P}(\boldsymbol{X}_r | \boldsymbol{X}_f, \boldsymbol{S}; \boldsymbol{\alpha}, \boldsymbol{\sigma}^2) = \prod_{j=1}^{d_r} \mathbb{P}(\boldsymbol{X}^{J_r^j} | \boldsymbol{X}^{J_p^j}, \boldsymbol{S}; \boldsymbol{\alpha}_j, \sigma_j^2). \tag{9}$$

Marginal model

We obtain for the distribution of $\{\boldsymbol{Y}|\boldsymbol{X}_f,\boldsymbol{S};\beta,\boldsymbol{\alpha},\sigma_Y^2,\boldsymbol{\sigma}^2\}$:

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}_f(\beta_f+\sum_{j=1}^{d_r}\beta_{J_r^j}\boldsymbol{\alpha}_j^*)+\sum_{j=1}^{d_r}\beta_{J_r^j}\boldsymbol{\varepsilon}_j+\boldsymbol{\varepsilon}_Y && (10)\\ &= \boldsymbol{X}_f\beta_f^*+\boldsymbol{\varepsilon}_Y^*, && (11)\end{aligned}$$

where $\boldsymbol{\alpha}_j^*\in\mathbb{R}^{d_f}$ with $(\boldsymbol{\alpha}_j^*)_{J_p^j}=\boldsymbol{\alpha}_j$ and $(\boldsymbol{\alpha}_j^*)_{J_f\setminus J_p^j}=\mathbf{0}$. We define $\boldsymbol{\alpha}^*\in\mathbb{R}^{(d_f\times d_r)}$ to use more compact notations:

$$\begin{aligned}\boldsymbol{X}_r &= \boldsymbol{X}_f\boldsymbol{\alpha}^*+\boldsymbol{\varepsilon}\\ \boldsymbol{Y} &= \boldsymbol{X}_f(\beta_f+\boldsymbol{\alpha}^*\beta_{\boldsymbol{r}})+\boldsymbol{\varepsilon}\beta_{\boldsymbol{r}}+\boldsymbol{\varepsilon}_Y && (12)\end{aligned}$$

Where $\boldsymbol{\varepsilon}$ is the $n\times d_r$ matrix whose columns are the $\boldsymbol{\varepsilon}_j$, the noises of the sub-regressions.

Plug-in model

$$\boldsymbol{\varepsilon}_Y^* = \boldsymbol{\varepsilon}\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}_Y. \tag{13}$$

Then the Best Linear Unbiased Estimator (BLUE) for $\boldsymbol{\beta}_r$ is given (MLE estimator) by:

$$\hat{\boldsymbol{\beta}}_r = (\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}_Y^*. \tag{14}$$

And we have the following estimators:

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= \boldsymbol{X}_r - \boldsymbol{X}_f\hat{\boldsymbol{\alpha}}^* \text{ and} \\ \hat{\boldsymbol{\varepsilon}}_Y^* &= \boldsymbol{Y} - \boldsymbol{X}_f\hat{\boldsymbol{\beta}}_f^* \end{aligned}$$

that we can use by plug-in.

Plug-in model

$$\hat{\boldsymbol{\beta}}_r^{\varepsilon} = (\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})^{-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}_Y^*$$

that depends on all covariates in $\boldsymbol{X}$ and relies on the estimated coefficients of sub-regressions $\hat{\boldsymbol{\alpha}}^*$ and on the estimate $\hat{\boldsymbol{\beta}}_f^*$ of the coefficients in the marginal model. Then we can estimate $\boldsymbol{Y}$ by:

$$\hat{\boldsymbol{Y}}_{plug-in} = \boldsymbol{X}_f\hat{\boldsymbol{\beta}}_f^* + \hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\beta}}_r^{\varepsilon}. \tag{15}$$

We can improve estimation of $\boldsymbol{\beta}_f$ (in terms of bias) by doing an additional identification step. We know that $\boldsymbol{\beta}_f^* = \boldsymbol{\beta}_f + \boldsymbol{\alpha}^*\boldsymbol{\beta}_r$ so we naturally define the following estimator:

$$\hat{\boldsymbol{\beta}}_f^{\varepsilon} = \hat{\boldsymbol{\beta}}_f^* - \hat{\boldsymbol{\alpha}}^*\hat{\boldsymbol{\beta}}_r^{\varepsilon}.$$

Marginal properties

biased

Plug-in properties

asymptotically unbiased

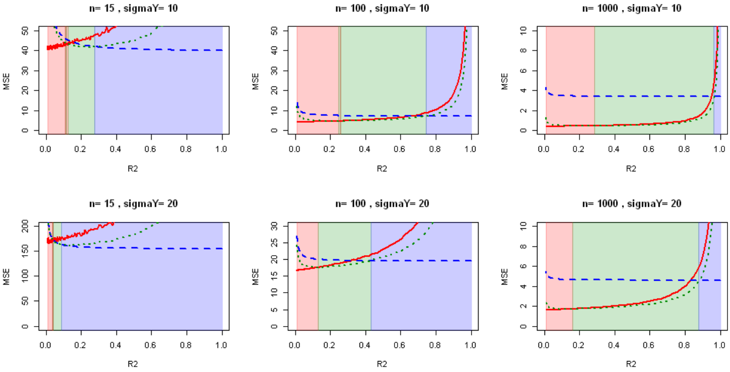

Figure: MSE on $\hat{\beta}$ of OLS (plain red) and `CorReg` marginal (blue dashed) and `CorReg` plug-in (green dotted) estimators for varying R^2 of the sub-regression, n and σ_Y. Results obtained on the running example with $d = 5$ covariates.

Lasso 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 better to find the true relevant covariates. Here $d = 3$ and $n = 1\,000$.
We define $\boldsymbol{X}^1, \boldsymbol{X}^2, \varepsilon_Y, \varepsilon_1 \quad i.i.d. \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ and then

$$\begin{aligned} \boldsymbol{X}^3 &= \frac{2}{3}\boldsymbol{X}^1 + \frac{2}{3}\boldsymbol{X}^2 + \frac{1}{3}\varepsilon_1 \text{ and} \\ \boldsymbol{Y} &= 2\boldsymbol{X}^1 + 3\boldsymbol{X}^2 + \varepsilon_Y. \end{aligned}$$

Lasso Consistency

True $\boldsymbol{S}$ was found 991 times on 1 000 tries.

	Classical LASSO	CorReg marginal + LASSO	CorReg full plug-in + LASSO
True $\boldsymbol{S}$	1.003303 (0.046)	**1.002273** (0.046)	**1.002812** (0.046)
$\hat{\boldsymbol{S}}$	1.003303 (0.046)	1.017622 (0.17)	**1.002812** (0.046)

Table: MSE observed on a validation sample (1 000 individuals) and their standard deviation (between brackets).

We look at the consistency that is the real stake:

	Classical LASSO	CorReg marginal + LASSO	CorReg full plug-in + LASSO
True $\boldsymbol{S}$	0	**1000**	**835**
$\hat{\boldsymbol{S}}$	0	**991**	**829**

Table: Number of consistent models found on 1 000 tries.

-

- Modèle génératif complet avec dépendances

- Explosion des mélanges

- SEM avec Gibbs

-

Bic pondéré

-

Résultats pourris

- Excel, fonctions graphiques, arbres de décision

Cule, E. (2014).
ridge: Ridge Regression with automatic selection of the penalty parameter.
R package version 2.1-3.

Cule, E. and De Iorio, M. (2013).
Ridge regression in prediction problems: automatic choice of the ridge parameter.
Genetic epidemiology, 37(7):704–714.

Dodge, Y. and Rousson, V. (2004).
Analyse de régression appliquée: manuel et exercices corrigés (coll. eco sup,).
Recherche, 67:02.

Er, M. J., Shao, Z., and Wang, N. (2013).
A systematic method to guide the choice of ridge parameter in ridge extreme learning machine.