Regression for correlated variables : Application in steel industry

Clément Théry [1]

[1] *ArcelorMittal Dunkerque, Inria Lille, Universit de Lille 1, clement.thery@arcelormittal.com*

Résumé. La régression linéaire suppose en général l'usage de variables explicatives indépendantes. Les variables présentes dans les bases de données d'origine industrielle sont souvent très fortement corrélées (de par le process, diverses lois physiques, etc). Le modèle génératif proposé consiste expliciter les corrélations présentes sous la forme d'une de sous-régressions linéaires. La structure est ensuite utilisée pour obtenir un modèle libéré des corrélations, facilement interprétable et compatible avec les méthodes de sélection de variables. La structure de corrélations est déterminée l'aide d'un algorithme de type MCMC. Un package R (CorReg) permet la mise en oeuvre de cette méthode.

Mots-clés. Régression, corrélations, industrie, sélection de variables, modèles génératifs, SEM (Structural Equation Model) ...

Abstract. Linear regression generally suppose independence between the covariates. Datasets found in industrial context often contains many highly correlated covariates (due to the process, physcial laws, etc). The proposed generative model consists in explicit modeling of the correlations with a structure of sub-regressions between the covariates. This structure is then used to obtain a model with independent covariates, easily interpreted, and compatible with any variable selection method. The structure of correlations is found with an MCMC algorithm. A R package (CorReg) implements this new method.

Keywords. Regression, correlations, industry, variable selection, generative models, Structural Equation Model ...

1 Le contexte

La rgression linaire classique suppose l'indpendance des covariables. Les corrlations sont problmatiques et posent des problmes.

$$Y = XA + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

$$Var(\hat{A}|X) = \sigma^2 (X'X)^{-1} \text{ explose si les colonnes de x sont linairement corrles} \tag{2}$$

2 Le modle gnratif

We have a p correlated covariates X to explain a response variabe Y. Let Z be the adjacency matrix that defines which covariates is depending on which others. That is $Z_{i,j} = \mathbf{1}_{(X^j depends\ on X^i)}$. We consider dependencies in a generative point of view ("depends on" means "is generated according to") so Z is not symmetric and has no cycles.

We can describe the structure Z by $S = (p_2, I_2, p_1, I_1)$ defined by :

$$
\begin{aligned}
p_2 &= \sum_{j=1}^{p} \mathbf{1}_{(\exists i, Z_{i,j} \neq 0)} \text{ the number of sub-regressions} && (3)\\
I_2 &= (I_2^1, \ldots, I_2^{p_2}) \text{ vector of the indices of the dependent covariates} && (4)\\
I_1 &= (I_1^1, \ldots, I_1^{p_2}) \text{ with} && (5)\\
I_1^j &= \{i | Z_{i,j} = 1\} \text{ indices of the covariates explaining } X^j && (6)\\
p_1 &= (p_1^1, \ldots, p_1^{p_2}) \text{ where } p_1^j = \sharp I_1^j && (7)
\end{aligned}
$$

We suppose $I_1 \cap I_2 = \emptyset$, *i.e.* dependent variables don't explain other variables in X.

We note $I_2^c = \{1, \ldots, p\} \setminus I_2$ Then our generative model can be written :

$$
\begin{aligned}
Y_{|X,S} = Y_{|X} &= XA + \varepsilon_Y = X^{I_2^c} A_{I_2^c} + X^{I_2} A_{I_2} + \varepsilon_Y \text{ with } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) && (8)\\
\forall j \in I_2 :\ X^j_{|X^{I_1^j},S} &= X^{I_1^j} B^j_{I_1^j} + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) && (9)\\
\forall j \notin I_2 :\ X^j &= f(\theta_j) \text{ free law} && (10)
\end{aligned}
$$

Where $B^j_{I_1^j}$ is the p_1^j-sized vector of the coefficients of the subregression.

We note that (8) and (9) also give :

$$
\begin{aligned}
Y &= X^{I_2^c}(A_{I_2^c} + \sum_{j \in I_2} B^j_{I_1} A_j) + \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y && (11)\\
&= X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} = X\tilde{A} + \tilde{\varepsilon} && (12)\\
\text{where } \tilde{A}_{I_2} &= 0 && (13)\\
\tilde{A}_{I_2^c} &= A_{I_2^c} + \sum_{j \in I_2} B^j_{I_1} A_j && (14)\\
\tilde{\varepsilon} &= \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y && (15)
\end{aligned}
$$

3 Estimateur

Classical methods like Ordinary Least Squares (OLS) estimate $Y|X$ and obtain (Maximum Likelihood Estimation):

$$\hat{A} = (X'X)^{-1}X'Y \text{ (ill-conditoned matrix to inverse)} \quad (16)$$

With following properties :

$$E[\hat{A}|X] = A \quad (17)$$
$$Var[\hat{A}|X] = \sigma_Y^2 (X'X)^{-1} \quad (18)$$

And when correlations are strong, the matrix to invert is ill-conditioned and the variance explodes.

Our idea is to reduce the variance so we explain Y only with X^{I_1} knowing (9) and (12)

$$Y = X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} \quad (19)$$

So the new estimator simply is :

$$\hat{\tilde{A}}_{I_2^c} = (X'_{I_2^c} X^{I_2^c})^{-1} X'_{I_2^c} Y \quad (20)$$
$$\hat{\tilde{A}}_{I_2} = 0 \quad (21)$$

and we get the following properties :

$$E[\hat{\tilde{A}}|X] = \tilde{A} \quad (22)$$
$$Var[\hat{\tilde{A}}_{I_2^c}|X] = (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 A_j^2)(X'_{I_2^c} X^{I_2^c})^{-1} \quad (23)$$
$$Var[\hat{\tilde{A}}_{I_2}|X] = 0 \quad (24)$$

We see that the variance is reduced (no correlations and smaller matrix give better conditioning) for small values of σ_j *i.e.* strong correlations.

Both classical and our new estimators of Y are unbiased (true model)[6].

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. Therefore we made some simulations to compare both methods (see the end of this paper). This new model is reduced even without variable selection and is just a linear regression so every method for variable selection in linear regression can be used. Hence we hope to obtain a parsimonious model.

The explicit structure between the covariates helps to understand the model and the complex link between the covariate and the response variable so we call this model explicative.

When we use a variable selection method on it we obtain two kinds of 0 :

1. Because of the structure we coerce $\hat{\tilde{A}}^{I_2} = 0$. This kind of zero means redundant information but the covariate can be correlated with the response variable. So we don't have the grouping effect (so we are more parsimonious) and we don't suffer from false interpretation (LASSO would).

2. Variable selection methods can lead to get some exact zeros in $\hat{\tilde{A}}^{I_1}$. This kind of zero means that implied covariate has no significant effect on the response variable. And because variables in X^{I_1} are orthogonal, we know that it is not misleading interpretation due to correlations.

4 Recherche de structure

All our work is based on a linear structure between the covariates. Let's define $\mathcal{S}$ and $\mathcal{Z}$ the ensemble of feasible structures and the ensemble of corresponding adjacency matrices. $Z \in \mathcal{Z} \Leftrightarrow$:

- Z is binary
- $ZZ = 0$ (Z is not crossed). Equivalent to $I_1 \cap I_2 = \emptyset$.

So $\mathcal{Z}$ is just the set of the binary square nilpotent matrices of size p. Z is an adjacency matrix and we know [1] that Z^p shows the number of paths of length p (linking $p+1$ vertices). So we suppose that Z is nilpotent, meaning it does not contain any non-trivial path. This strong hypothesis also strongly reduces the size of $\mathcal{Z}$.

4.1 Comparaison des structures

On utilise le Bayesian Information criterion (BIC) [4]. But BIC tends to give too complex structures because we test a great range of models. Thus we choose to penalise the complexity a bit more with specific a priori laws(uniform laws for the number of subregression

and the complexity of each subregression instead of uniform law on S) :

$$
\begin{aligned}
P(S) &= P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2) && (25)\\
P(I_1|p_1, I_2, p_2) &= \prod_{j=1}^{p_2} P(I_1^j|p_1^j, I_2, p_2) && (26)\\
P(I_1^j|p_1^j, I_2, p_2) &= \left(\begin{array}{c} p-p_2 \\ p_1^j \end{array} \right)^{-1} = \frac{p_1^j!(p-p_2-p_1^j)!}{(p-p_2)!} && (27)\\
P(p_1|I_2, p_2) &= \prod_{j=1}^{p_2} P(p_1^j|I_2, p_2) && (28)\\
P(p_1^j|I_2, p_2) &= \frac{1}{p-p_2} && (29)\\
P(I_2|p_2) &= \left(\begin{array}{c} p \\ p_2 \end{array} \right)^{-1} = \frac{p_2!(p-p_2)!}{p!} && (30)\\
P(p_2) &= \frac{1}{p_2} && (31)\\
P(S) &= \left(\prod_{j=1}^{p_2} \left(\begin{array}{c} p-p_2 \\ p_1^j \end{array} \right)^{-1} \right) \left(\frac{1}{p-p_2} \right)^{p_2} \frac{p_2!(p-p_2)!}{p!} \frac{1}{p_2} && (32)\\
\ln P(S) &= -\sum_{j=1}^{p_2} \ln \left(\begin{array}{c} p-p_2 \\ p_1^j \end{array} \right) - p_2 \ln(p-p_2) - \ln \left(\begin{array}{c} p \\ p_2 \end{array} \right) - \ln(p_2) && (33)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
P(S|X) &\propto P(X|S)P(S) && (34)\\
\ln(P(S|X)) &= \ln(P(X|S)) + \ln(P(S)) + cste && (35)\\
BIC^* &= BIC + \ln(P(S)) && (36)
\end{aligned}
$$

It increases penalty on complexity for $p_2 < \frac{p}{2}$ thus in the following we will use $BIC*$ under this hypothesis (that becomes a constraint in the MCMC).

$$BIC(X|S) = \sum_{j=1}^{p} BIC(X^j|S) \quad (37)$$

Where

$$BIC(X^j|S) = -2\mathcal{L}_{|S}(X^j, \theta_j) + K_j \log(n) \quad (38)$$

Where K_j is the number of parameters to estimate. We will now use the following notation : $BIC(S) = BIC(X|S)$ If we have some hypothesis on the distribution of some variables (exponentially distributed for example) we can compute corresponding BIC separately and then improve the efficiency of the algorithm (it will find a structure only if it is really relevant).

4.2 The Markov chain

First we find that S is completely described with I_1 :

$$\begin{aligned} I_2 &= \{j|\sharp I_1^j > 0\} && (39)\\ p_2 &= \sharp I_2 && (40)\\ \forall j p_1^j &= \sharp I_1^j && (41) \end{aligned}$$

So we will only describe the variations in I_1 at each step and other parts of S will follow according to the previous definition. for each step, starting from $S \in \mathcal{S}$ we define a neighbourhood $\mathcal{V}_{S,j}$ with $j \sim \mathcal{U}(\{1,\dots,p\})$ like this :

$$\begin{aligned} \mathcal{V}_{S,j} = \{\tilde{S} \in \mathcal{S} \quad | \quad & \exists! i, \tilde{Z}_{i,j} = 1 - Z_{i,j}, \text{ and } \forall (k,l) \neq (i,j) : \\ & \tilde{Z}_{j,l} = (1 - \tilde{Z}_{i,j}) Z_{j,l} \text{ (row-wise relaxation)}, \\ & \tilde{Z}_{k,i} = (1 - \tilde{Z}_{i,j}) Z_{k,i} \text{ (column-wise relaxation)} \\ & \tilde{Z}_{k,l} = Z_{k,l}\} \cup \{Z\} \end{aligned}$$

We have $|\mathcal{V}_{Z,j}| = p$ but some other constraints can be added on the definition of $\mathcal{Z}$ and will consequently modify the size of the neighbourhood (for example a maximum complexity for the subregressions or the whole structure). The algorithm follows a time-homogeneous markov chain whose transition matrix $\mathcal{P}$ has $|\mathcal{Z}|$ rows and columns (combinatory so we'll just compute the probabilities when we need them). And $\mathcal{Z}$ is the finite state space.We want

$$\mathcal{P}(Z, \tilde{Z}) = \mathbf{1}_{[\exists j, \tilde{Z} \in \mathcal{V}_{Z,j}]} P(\tilde{Z}|X) \tag{42}$$

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [3] : π and every rows of $\lim_{k\to\infty} \mathcal{P}^k = W$ equals π. With $\forall Z \in \mathcal{Z}$:

$$\begin{aligned} 0 \leq \quad \pi(Z) \quad &\leq 1 && (43)\\ \sum_{Z \in \mathcal{Z}} \pi(Z) \quad &= \quad 1 && (44)\\ \pi(Z) \quad &= \quad \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) && (45)\\ &= \quad \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) && (46) \end{aligned}$$

$$P(Z|X) = P(X|Z)P(Z) \propto P(X|Z) \tag{47}$$

We make a first approximation :

$$P(X|Z) \approx exp(BIC(Z)) \tag{48}$$

We define [4], :

$$q(\tilde{Z}, \mathcal{V}_{Z,j}) = \mathbf{1}_{\{\tilde{Z} \in \mathcal{V}_{Z,j}\}} \frac{exp(\frac{-1}{2}\Delta BIC(\tilde{Z}, \mathcal{V}_{Z,j}))}{\sum_{Z_l \in \mathcal{V}_{Z,j}} exp(\frac{-1}{2}\Delta BIC(Z_l, \mathcal{V}_{Z,j}))} \tag{49}$$

where $\Delta BIC(Z, \mathcal{V}_{Z,j}) = BIC(Z) - \min\{BIC(\tilde{Z}) | \tilde{Z} \in \mathcal{V}_{Z,j}\}$ is the gap between a structure and the worst structure in the neighbourhood in terms of BIC.
And then we can note $\forall (Z, \tilde{Z}) \in \mathcal{Z}^2$:

$$\mathcal{P}(Z, \tilde{Z}) = \frac{1}{p} \sum_{j=1}^{p} q(\tilde{Z}, \mathcal{V}_{Z,j})$$

The output will be the best structure seen in terms of BIC. If we have some knowledge about some sub-regressions (physical models for example) we can add them in the found structure. So the model is really expert-friendly. Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [3] : π and every rows of $\lim_{k\to\infty} \mathcal{P}^k = W$ equals π. With $\forall Z \in \mathcal{Z}$:

$$0 \leq \pi(Z) \leq 1 \tag{50}$$

$$\sum_{Z \in \mathcal{Z}} \pi(Z) = 1 \tag{51}$$

$$\pi(Z) = \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \mathcal{P}(\tilde{Z}, Z) \tag{52}$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^{p} q(Z, \mathcal{V}_{\tilde{Z},j}) \tag{53}$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{Z \in \mathcal{V}_{\tilde{Z},j}\}} \frac{exp(\frac{-1}{2}\Delta BIC(Z, \mathcal{V}_{\tilde{Z},j}))}{\sum_{Z_l \in \mathcal{V}_{\tilde{Z},j}} exp(\frac{-1}{2}\Delta BIC(Z_l, \mathcal{V}_{\tilde{Z},j}))} \tag{54}$$

$$= \sum_{\tilde{Z} \in \mathcal{Z}} \pi(\tilde{Z}) \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{Z \in \mathcal{V}_{\tilde{Z},j}\}} \frac{exp(\frac{-1}{2}BIC(Z))}{\sum_{Z_l \in \mathcal{V}_{\tilde{Z},j}} exp(\frac{-1}{2}BIC(Z_l))} \tag{55}$$

The initial structure is based on a first warming algorithm taking the correlations into account. Ones are randomly placed into Z, weighted by the absolute value of the correlations. Then this structure is reduced by the hadamard product with the binary matrix obtained by Graphical Lasso[2].

5 Rsultats

6 Conclusion et perspectives

CorReg est fonctionnel et disponible Besoin d'largir la gestion des valeurs manquantes trs prsentes dans l'industrie

7 Exemple de références bibliographiques

La nécessité de produire des résumés clairs et bien référencés a été démontrée par Achin et Quidont (2000). Le récent article de Noteur (2003) met en évidence ...

Bibliographie

[5]

References

[1] Norman Biggs. *Algebraic graph theory*. Cambridge University Press, 1993.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[3] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1997.

[4] É. Lebarbier and T. Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–57, 2006.

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

[6] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

recopier dans le bon ordre comme demandé ci-dessous. [1] Auteurs (année), Titre, revue, localisation. [2] Achin, M. et Quidont, C. (2000), *Théorie des Catalogues*, Editions du Soleil, Montpellier. [3] Noteur, U. N. (2003), Sur l'intérêt des résumés, *Revue des Organisateurs de Congrès*, 34, 67–89.