

REGRESSION FOR CORRELATED VARIABLES : APPLICATION IN STEEL INDUSTRY

Clément Théry ¹

¹ *ArcelorMittal Dunkerque, Inria Lille, Université 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 à 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 à 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, physical 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 régression linéaire classique suppose l’indépendance des covariables. Les corrélations posent des problèmes.

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

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

2 Le modèle génératif

On dispose de p variables X fortement corrélées pour expliquer une variable réponse Y . On rend explicite les corrélations au sein de X sous la forme d'une structure de sous-régressions linéaires $S = (p_2, I_2, p_1, I_1)$ définie ainsi :

$$I_1 = (I_1^1, \dots, I_1^{p_2}) \text{ avec} \quad (3)$$

$$I_1^j = \{i | Z_{i,j} = 1\} \text{ indices des covariables qui expliquent } X^j \quad (4)$$

$$I_2 = \{j | \#I_1^j > 0\} \text{ indices des variables dépendantes} \quad (5)$$

$$p_2 = \#I_2 \quad (6)$$

$$\forall 1 \leq j \leq p : p_1^j = \#I_1^j \quad (7)$$

On suppose $I_1 \cap I_2 = \emptyset$, *i.e.* Les variables dépendantes dans X n'en expliquent pas d'autres.

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

$$Y_{|X,S} = Y_{|X} = XA + \varepsilon_Y = X^{I_2^c} A_{I_2^c} + X^{I_2} A_{I_2} + \varepsilon_Y \text{ with } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) \quad (8)$$

$$\forall j \in I_2 : X_{|X^{I_1^j}, S}^j = X^{I_1^j} B_{I_1^j}^j + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) \quad (9)$$

$$\forall j \notin I_2 : X^j = f(\theta_j) \text{ free law} \quad (10)$$

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

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

$$Y = X^{I_2^c} (A_{I_2^c} + \sum_{j \in I_2} B_{I_1^j}^j A_j) + \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y \quad (11)$$

$$= X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} = X \tilde{A} + \tilde{\varepsilon} \quad (12)$$

$$\text{where } \tilde{A}_{I_2} = 0 \quad (13)$$

$$\tilde{A}_{I_2^c} = A_{I_2^c} + \sum_{j \in I_2} B_{I_1^j}^j A_j \quad (14)$$

3 Estimateur

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

$$\hat{A} = (X'X)^{-1} X'Y \text{ (ill-conditioned matrix to inverse)} \quad (15)$$

With following properties :

$$E[\hat{A}|X] = A \quad (16)$$

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

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)

$$Y = X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} \quad (18)$$

So the new estimator simply is :

$$\hat{\tilde{A}}_{I_2^c} = (X_{I_2^c}' X^{I_2^c})^{-1} X_{I_2^c}' Y \quad (19)$$

$$\hat{\tilde{A}}_{I_2} = 0 \quad (20)$$

and we get the following properties :

$$E[\hat{\tilde{A}}|X] = \tilde{A} \quad (21)$$

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

$$Var[\hat{\tilde{A}}_{I_2}|X] = 0 \quad (23)$$

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)[5].

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.

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

4.1 Comparaison des structures

On utilise le Bayesian Information criterion (BIC) [3]. But BIC tends to give too complex structures because we test a great range of models. Thus we choose to penalise the com-

plexity 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) :

$$P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2) \quad (24)$$

$$P(I_1|p_1, I_2, p_2) = \prod_{j=1}^{p_2} P(I_1^j|p_1^j, I_2, p_2) \quad (25)$$

$$P(I_1^j|p_1^j, I_2, p_2) = \binom{p-p_2}{p_1^j}^{-1} = \frac{p_1^j!(p-p_2-p_1^j)!}{(p-p_2)!} \quad (26)$$

$$P(p_1|I_2, p_2) = \prod_{j=1}^{p_2} P(p_1^j|I_2, p_2) \quad (27)$$

$$P(p_1^j|I_2, p_2) = \frac{1}{p-p_2} \quad (28)$$

$$P(I_2|p_2) = \binom{p}{p_2}^{-1} = \frac{p_2!(p-p_2)!}{p!} \quad (29)$$

$$P(p_2) = \frac{1}{p_2} \quad (30)$$

$$P(S) = \left(\prod_{j=1}^{p_2} \binom{p-p_2}{p_1^j}^{-1} \right) \left(\frac{1}{p-p_2} \right)^{p_2} \frac{p_2!(p-p_2)!}{p!} \frac{1}{p_2} \quad (31)$$

$$\ln P(S) = - \sum_{j=1}^{p_2} \ln \binom{p-p_2}{p_1^j} - p_2 \ln(p-p_2) - \ln \binom{p}{p_2} - \ln(p_2) \quad (32)$$

Then we have

$$P(S|X) \propto P(X|S)P(S) \quad (33)$$

$$\ln(P(S|X)) = \ln(P(X|S)) + \ln(P(S)) + cste \quad (34)$$

$$BIC^* = BIC + \ln(P(S)) \quad (35)$$

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).

4.2 The Markov chain

First we remember that S is completely described with I_1 : 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 :

$$\mathcal{V}_{S,j} = \{S^{(i,j)} | 1 \leq i \leq p\} \cup \{S\} \quad (36)$$

With $S^{(i,j)}$ defined by the following algorithm :

- if $i \notin I_i^j$ (add):
 - $I_1^j = I_1^j \cup \{i\}$
 - $I_1^i = \emptyset$ (explicative variables can't depend on others : column-wise relaxation)
 - $I_1 = I_1 \setminus \{j\}$ (dependent variables can't explain others : row-wise relaxation)
- if $i \in I_1^j$ (remove): $I_1^j = I_1^j \setminus \{i\}$

And then :

We have $|\mathcal{V}_{S,j}| = p$ but some other constraints can be added on the definition of \mathcal{S} and will consequently modify the size of the neighbourhood (for example a maximum complexity for the subregressions or the whole structure, a maximum number of sub-regressions, etc). CorReg allows to modify this neighbourhood to better fit users constraints.

We make a first approximation (??) :

$$P(S|X) \approx \exp(BIC^*(S)) \quad (37)$$

We define [3], :

$$q(\tilde{S}, \mathcal{V}_{S,j}) = \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(\frac{-1}{2} \Delta BIC(\tilde{S}, \mathcal{V}_{S,j}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(\frac{-1}{2} \Delta BIC(S_l, \mathcal{V}_{S,j}))} \quad (38)$$

Where $\Delta BIC(S, \mathcal{V}_{S,j}) = BIC(S) - \min\{BIC(\tilde{S}) | \tilde{S} \in \mathcal{V}_{S,j}\}$ is the gap between a structure and the worst structure in the neighbourhood in terms of BIC.

And then we can note $\forall (S, \tilde{S}) \in \mathcal{S}^2$:

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

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.

4.2.1 initialisation

L'initialisation peut par exemple utiliser la matrice des corrélations et/ou la méthode du Graphical Lasso[1]. La grande dimension de l'espace parcouru rend préférable l'utilisation de multiples chanes courtes plutt qu'une seule trs longue (pour un temps de calcul égal). Accessoirement, les multiples chanes permettent de paralléliser la recherche, ce qui peut tre trs appréciable.

4.2.2 properties

The algorithm follows a time-homogeneous markov chain whose transition matrix \mathcal{P} has $|\mathcal{S}|$ rows and columns (combinatory so we'll just compute the probabilities when we need them). And \mathcal{S} is a finite state space.

We want

$$\mathcal{P}(S, \tilde{S}) = \mathbf{1}_{[\exists j, \tilde{S} \in \mathcal{V}_{S,j}]} P(\tilde{S}|X) \quad (39)$$

Because the walk follows a regular and thus ergodic markov chain with a finite state space, it has exactly one stationary distribution [2] : π and every rows of $\lim_{k \rightarrow \infty} \mathcal{P}^k = W$ equals π .

5 Résultats

6 Conclusion et perspectives

CorReg est fonctionnel et disponible. Besoin d'élargir la gestion des valeurs manquantes très présentes 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

[4]

References

- [1] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [2] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 1997.
- [3] É. Lebarbier and T. Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–57, 2006.
- [4] D.W. Marquardt and R.D. Snee. Ridge regression in practice. *American Statistician*, pages 3–20, 1975.

[5] 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 Organismes de Congrès*, 34, 67–89.