

CORREG : RÉGRESSION SUR VARIABLES CORRÉLÉES ET APPLICATION À L'INDUSTRIE SIDÉRURGIQUE

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 structure 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 en effet des problèmes, tant au niveau de l'interprétation qu'en termes de variance des estimateurs.

$$\begin{aligned} Y &= XA + \varepsilon \quad \text{avec } \varepsilon \sim \mathcal{N}(0, \sigma^2) \\ \text{Var}(\hat{A}|X) &= \sigma^2(X'X)^{-1} \text{ explose si les colonnes de } x \text{ sont linéairement corrélées} \end{aligned} \quad (1)$$
$$(2)$$

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 :

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

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

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

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

$$p_1 = (p_1^1, \dots, p_1^{p_2}) \text{ avec } p_1^j = \#I_1^j \quad (7)$$

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

On note $I_2^c = \{1, \dots, p\} \setminus I_2$ Le modèle génératif s'écrit alors :

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

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

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

Où $B_{I_1^j}^j$ est le vecteur de taille p_1^j des coefficients de la sous-régression de X^j .

On remarque que (8) et (9) impliquent :

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

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

3 Estimateur

Les Moindres Carrés Ordinaires (MCO) donnent (maximum de vraisemblance):

$$\hat{A} = (X'X)^{-1} X'Y \text{ (matrice à inverser mal conditionnée)} \quad (13)$$

avec les propriétés suivantes :

$$E[\hat{A}|X] = A \quad (14)$$

$$\text{Var}[\hat{A}|X] = \sigma_Y^2 (X'X)^{-1} \quad (15)$$

La variance de l'estimateur explose quand les corrélations sont fortes (matrice quasi-singulière).

CorReg réduit la variance de en estimant Y seulement avec X^{I_1} , sachant (9) et (12).

$$Y = X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} \quad (16)$$

Ainsi, l'estimateur devient :

$$\hat{\tilde{A}}_{I_2^c} = (X_{I_2^c}' X^{I_2^c})^{-1} X_{I_2^c}' Y \quad (17)$$

$$\hat{\tilde{A}}_{I_2} = 0 \quad (18)$$

avec les propriétés suivantes :

$$E[\hat{\tilde{A}}|X] = \tilde{A} \quad (19)$$

$$\text{Var}[\hat{\tilde{A}}_{I_2^c}|X] = (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 A_j^2)(X_{I_2^c}' X^{I_2^c})^{-1} \quad (20)$$

$$\text{Var}[\hat{\tilde{A}}_{I_2}|X] = 0 \quad (21)$$

La variance est réduite(retrait des corrélations et réduction de la dimension améliorent drastiquement le conditionnement) pour les faibles valeurs de σ_j *i.e.* les fortes corrélations.

Le modèle complet et le nôtre prédisent tous les deux Y sans biais (vrai modèle)[3]. Ce nouveau modèle est de dimension réduite et consiste en une régression linéaire classique qui peut donc bénéficier des outils de sélection de variables au même titre que le modèle complet.

La structure explicite entre les variables permet de mieux comprendre les phénomènes en jeu et la parsimonie du modèle facilite son interprétation.

En ajoutant une étape de sélection de variable on obtient deux types de 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 [4]).
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

On va s'appuyer sur la vraisemblance pénalisée de la structure à la manière du critère BIC [2].

$$P(S|X) \propto P(X|S)P(S) \quad (22)$$

$$\ln(P(S|X)) = \ln(P(X|S)) + \ln(P(S)) + cste \quad (23)$$

$$BIC^* = BIC + \ln(P(S)) \quad (24)$$

Pour éviter une surcomplexité de la structure trouvée, on peut alors faire des hypothèses a priori sur $P(S)$. Par exemple, au lieu de supposer l'équiprobabilité pour tous les S , on peut supposer l'équiprobabilité des p_2 et p_1^j , ce qui vient pénaliser davantage la complexité sous l'hypothèse $p_2 < \frac{p}{2}$ (qui devient alors une contrainte supplémentaire dans l'algorithme de recherche). On a

$$P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2) \quad (25)$$

S est entièrement défini à partir de I_1 donc on se contente ici de modifier I_1 . A chaque étape, pour $S \in \mathcal{S}$ on définit un voisinage $\mathcal{V}_{S,j}$ avec $j \sim \mathcal{U}(\{1, \dots, p\})$:

$$\mathcal{V}_{S,j} = \{S^{(i,j)} | 1 \leq i \leq p\} \cup \{S\} \quad (26)$$

avec $S^{(i,j)}$ obtenu selon l'algorithme :

- Si $i \notin I_1^j$ (ajout):
 - $I_1^j := I_1^j \cup \{i\}$, et pour garder $I_1 \cap I_2 = \emptyset$:
 - $I_1^i := \emptyset$ et $I_1 := I_1 \setminus \{j\}$
- Sinon ($i \in I_1^j$ (suppression)): $I_1^j = I_1^j \setminus \{i\}$

On a donc p candidats à chaque étape. Mais le package CorReg permet à l'utilisateur de modifier ce voisinage.

We make a first approximation (24) :

$$P(S|X) \approx \exp(BIC^*(S)) \quad (27)$$

$$q(\tilde{S}, \mathcal{V}_{S,j}) = \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(\frac{-1}{2}\Delta BIC(\tilde{S}, \mathcal{V}_{S,j}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(\frac{-1}{2}\Delta BIC(S_l, \mathcal{V}_{S,j}))} \quad (28)$$

Where $\Delta BIC(S, \mathcal{V}_{S,j}) = BIC(S) - \min\{BIC(\tilde{S}) | \tilde{S} \in \mathcal{V}_{S,j}\}$.

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

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

La chaîne de Markov ainsi constituée est ergodique dans un espace d'états finis et possède une unique loi stationnaire. Le résultat obtenu est la meilleure structure rencontrée en termes de BIC^* (vraisemblance pénalisée). Certaines parties de S peuvent être contraintes pour insérer par exemple des modèles physiques. De ce fait, la méthode permet de tenir compte d'éventuels modèles experts.

L'initialisation peut se faire en utilisant 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 chaînes courtes plutôt qu'une seule très longue (pour un temps de calcul égal). Accessoirement, les multiples chaînes permettent de paralléliser la recherche, ce qui peut être très appréciable.

5 Résultats

Les données industrielles sont fortement corrélées de manière naturelle : Largeur et poids d'une brame d'acier ($\rho = 0.905$), Température avant et après un outil ($\rho = 0.983$), rugosité des deux faces du produit ($\rho = 0.919$), Moyenne et maximum d'une courbe ($\rho = 0.911$).

Certaines des sous-régressions obtenues par CorReg ont une interprétation physique:

- Moyenne = f (Min , Max , Sigma) pour des données courbes
- Largeur du produit= f (débit de fonte , vitesse de la coulée continue) Vrai modèle physique (non linéaire) :

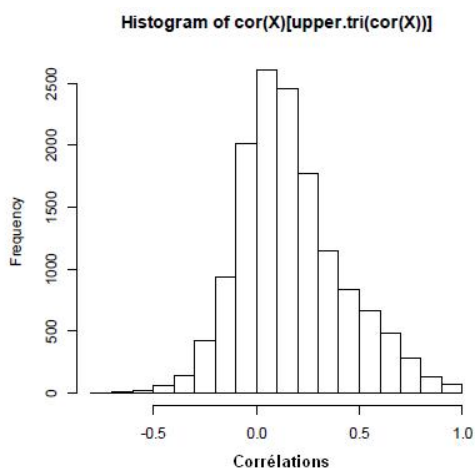
$$\text{Largeur} = \frac{\text{débit}}{\text{vitesse} \times \text{épaisseur}} \text{ (Mais dans ce cas précis l'épaisseur est constante)}$$

D'autres sous-régressions traduisent des modèles physiques (modèles ballistiques, etc) utilisés pour réguler le process.

Exemple de régression finale :

6 Conclusion et perspectives

CorReg est fonctionnel et disponible. L'outil a d'ores et déjà montré son efficacité sur de vraies problématiques de régression en entreprise. La force de CorReg est la grande interprétabilité du modèle proposé, qui est constitué de plusieurs modèles simples (parsimonieux) et facilement accessibles aux non statisticiens (régressions linéaires)



	MSE	nombre de régresseurs
LASSO (lars)	0.80	54
CorReg (et lars)	0.53	24

Figure 1: résultats obtenus sur données réelles : $n = 117$ et $p = 168$. À gauche les corrélations entre variables explicatives, à droite les MSE et la complexité (nombre de variables explicatives) des modèles obtenus pour expliquer Y

tout en luttant efficacement contre les problématiques de corrélations, omniprésentes dans l’industrie. On note néanmoins le besoin d’élargir le champ d’application à la gestion des valeurs manquantes, très présentes dans l’industrie. Cet aspect est envisagé sérieusement pour la prochaine version de CorReg.

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

References

- [1] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [2] É. Lebarbier and T. Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS*, 147(1):39–57, 2006.
- [3] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.
- [4] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006.

à 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.