CorReg : Régression sur variables Corrélées et Application à l'industrie Sidérurgique

Clément Théry[1] & Christophe Biernacki[2] & Gaétan Loridant[3]

[1] *ArcelorMittal, Université Lille 1, Inria, CNRS, clement.thery@inria.fr*
[2] *Université Lille 1, Inria, CNRS, christophe.biernacki@math.univ-lille1.fr*
[3] *Etudes Industrielles ArcelorMittal Dunkerque, gaetan.loridant@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 ici 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 parcimonieux 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, 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 reduced model with independent covariates, easily interpreted, and compatible with any variable selection method. The structure of correlations is found with an MCMC algorithm. An R package (CorReg) implements this new method.

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

1 Décorrélation par modèle génératif

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. La régression $Y = XA + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ donne un estimateur de variance $\mathrm{Var}\,(\hat{A}|X) = \sigma^2(X'X)^{-1}$ qui explose si les colonnes de X sont linéairement corrélées

On suppose le modèle génératif suivant :

$$
\begin{aligned}
Y_{|X,S} &= XA + \varepsilon_Y = X^{I_2^c} A_1 + X^{I_2} A_2 + \varepsilon_Y \text{ avec } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) && (1)\\
\forall j \in I_2 :\ X^j_{|X^{I_1^j},S} &= X^{I_1^j} B^j_{I_1^j} + \varepsilon_j \text{ avec } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) && (2)\\
\forall j \notin I_2 :\ X^j &= f(\theta_j) \text{ loi quelconque} && (3)
\end{aligned}
$$

Où $B^j_{I_1^j}$ est le vecteur de taille p_1^j des coefficients de la sous-régression en X^j, $I_1 = \{I_1^1, \ldots, I_1^p\}$, $I_2 = \{j | \sharp I_1^j > 0\}$. On suppose $I_1 \cap I_2 = \emptyset$, *i.e.* Les variables dépendantes dans X n'en expliquent pas d'autres. On note $p_2 = \sharp I_2$, $p_1 = (p_1^1, \ldots, p_1^{p_2})$ et $I_2^c = \{1, \ldots, p\} \setminus I_2$.

On a donc rendu explicites les corrélations au sein de X sous la forme d'une structure de sous-régressions linéaires $S = (p_2, I_2, p_1, I_1)$.

On remarque que (1) et (2) impliquent :

$$
\begin{aligned}
Y_{|X,S} &= X^{I_2^c}(A_{I_2^c} + \sum_{j \in I_2} B^j_{I_1} A_j) + \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y && (4)\\
&= X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} = X\tilde{A} + \tilde{\varepsilon} && (5)
\end{aligned}
$$

2 Estimateur

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

$$\hat{A} = (X'X)^{-1}X'Y \text{ (matrice à inverser mal conditionnée)} \quad (6)$$

avec les propriétés suivantes :

$$
\begin{aligned}
\mathrm{E}[\hat{\mathrm{A}}|\mathrm{X}] &= A && (7)\\
\mathrm{Var}\,[\hat{A}|X] &= \sigma_Y^2 (X'X)^{-1} && (8)
\end{aligned}
$$

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

CorReg réduit la variance de l'estimateur en estimant Y seulement à partir de X^{I_1}, sachant (2) et (5).

$$Y = X^{I_2^c} \tilde{A}_{I_2^c} + \tilde{\varepsilon} \quad (9)$$

Ainsi, l'estimateur devient :

$$
\begin{aligned}
\hat{\tilde{A}}_{I_2^c} &= (X'_{I_2^c} X^{I_2^c})^{-1} X'_{I_2^c} Y && (10)\\
\hat{\tilde{A}}_{I_2} &= 0 && (11)
\end{aligned}
$$

avec les propriétés suivantes :

$$\begin{aligned} \mathrm{E}[\hat{\tilde{\mathrm{A}}}|\mathrm{X}] &= \tilde{A} && (12)\\ \mathrm{Var}\,[\hat{\tilde{A}}_{I_2^c}|X] &= (\sigma_Y^2 + \sum_{j\in I_2} \sigma_j^2 A_j^2)(X'_{I_2^c} X^{I_2^c})^{-1} && (13)\\ \mathrm{Var}\,[\hat{\tilde{A}}_{I_2}|X] &= 0 && (14) \end{aligned}$$

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 parcimonie du modèle facilite son interprétation.

En ajoutant une étape de sélection de variable on obtient deux types de 0 :

1. La structure implique $\hat{\tilde{A}}^{I_2} = 0$. Ce type de 0 est à interpréter comme une redondance d'information, mais la variable associée n'est pas pour autant jugée indépendante de Y. On obtient un modèle parcimonieux sans effet groupe, mais sans erreur d'interprétation, contrairement au LASSO [4] car on connaît la signification de ce type de 0.

2. Une phase de sélection de variables peut également produire des coefficients nuls dans $\hat{\tilde{A}}^{I_1}$. Ces 0 sont alors des 0 d'indépendance vis-à-vis de la réponse qui s'interprètent donc classiquement. Et comme les variables de X^{I_1} sont orthogonales, on peut avoir confiance dans cette interprétation.

Le modèle obtenu est donc sans biais de prédiction, parcimonieux et consistant en interprétation.

3 Recherche de structure

Le choix de structure s'appuie sur BIC^*, vraisemblance pénalisée de la structure à la manière du critère BIC [2].

$$\begin{aligned} P(S|X) &\propto P(X|S)P(S) && (15)\\ \ln(P(S|X)) &= \ln(P(X|S)) + \ln(P(S)) + cste && (16)\\ BIC^* &= BIC + \ln(P(S)) && (17) \end{aligned}$$

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 compléxité sous l'hypothèse $p_2 < \frac{p}{2}$ (qui devient alors une contrainte supplémentaire dans l'algorithme de recherche). On a

$$P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2) \tag{18}$$

A chaque étape de l'algorithme MCMC, pour $S \in \mathcal{S}$ (ensemble des structures réalisables) on définit un voisinage $\mathcal{V}_S$ de p candidats , mais le package CorReg permet à l'utilisateur de modifier ce voisinage. S est entièrement défini à partir de I_1 donc on se contente ici de décrire les modifications dans I_1.

On fait l'approximation suivante :

$$P(S|X) \approx \exp(BIC^*(S)) \tag{19}$$

On définit alors

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) = \frac{1}{p} \sum_{j=1}^{p} q(\tilde{S}, \mathcal{V}_{S,j}) \tag{20}$$

$$\text{où } q(\tilde{S}, \mathcal{V}_{S,j}) = \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_{S,j}\}} \frac{\exp(\frac{-1}{2}\Delta BIC(\tilde{S}, \mathcal{V}_{S,j}))}{\sum_{S_l \in \mathcal{V}_{S,j}} \exp(\frac{-1}{2}\Delta BIC(S_l, \mathcal{V}_{S,j}))} \tag{21}$$

$$\text{avec } \Delta BIC(S, \mathcal{V}_{S,j}) = BIC(S) - \min\{BIC(\tilde{S}) | \tilde{S} \in \mathcal{V}_{S,j}\} \tag{22}$$

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).

L'intialisation 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 (pour un temps de calcul égal) l'utilisation de multiples chaînes courtes plutôt qu'une seule très longue (permet aussi la parallélisation).

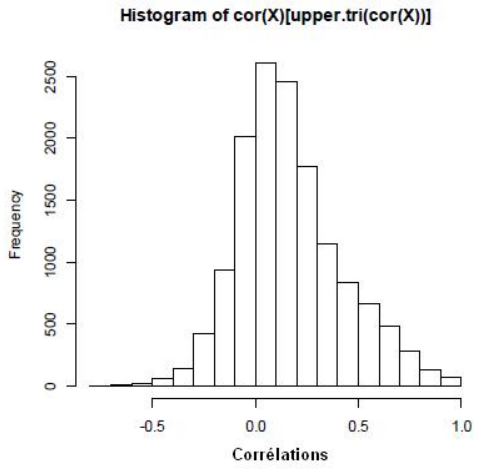
4 Résultats

Les données industrielles sont fortement corrélées de manière naturelle : largeur et poids d'une brame ($\rho = 0.905$), température avant et après un outil ($\rho = 0.983$), rugosité des deux faces du produit ($\rho = 0.919$), Moyenne et maximum d'une courbe ($\rho = 0.911$). Exemples de Sous-régressions obtenues par CorReg ayant 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) :
  Largeur = $\frac{\text{débit}}{\text{vitesse} \times \text{épaisseur}}$ (Mais dans ce cas précis l'épaisseur est constante)

D'autres sous-régressions traduisent des modèles physiques qui régulent le process...

Exemple de régression sur une variable réponse dans le cadre des données réelles :



	MSE	Variables retenues
LASSO (lars)	0.80	54
CorReg (et lars)	0.53	24

Figure 1: résultats obtenus sur données réelles : $n = 117$ et $p = 168$. l'erreur est réduite d'un tiers alors que la complexité du modèle est divisée par $2,5$.

5 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 (parcimonieux) et facilement accessibles aux non statisticiens (régressions linéaires) 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.

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.