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, 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 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.

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

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

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 :

$$
\begin{aligned}
I_1 &= (I_1^1, \ldots, I_1^{p_2}) \text{ avec} && (3)\\
I_1^j &= \{i | Z_{i,j} = 1\} \text{ indices des covariables qui expliquent } X^j && (4)\\
I_2 &= \{j | \sharp I_1^j > 0\} \text{ indices des variables dépendantes} && (5)\\
p_2 &= \sharp I_2 && (6)\\
p_1 &= (p_1^1, \ldots, p_1^{p_2}) \text{ avec } p_1^j = \sharp I_1^j && (7)
\end{aligned}
$$

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

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

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

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 :

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

3 Estimateur

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

$$\hat{A} = (X'X)^{-1}X'Y \text{ (ill-conditoned matrix to inverse)} \quad (13)$$

With following properties :

$$
\begin{aligned}
E[\hat{A}|X] &= A && (14)\\
Var[\hat{A}|X] &= \sigma_Y^2 (X'X)^{-1} && (15)
\end{aligned}
$$

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)

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

So the new estimator simply is :

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

$$\hat{\tilde{A}}_{I_2} = 0 \tag{18}$$

and we get the following properties :

$$E[\hat{\tilde{A}}|X] = \tilde{A} \tag{19}$$

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

$$Var[\hat{\tilde{A}}_{I_2}|X] = 0 \tag{21}$$

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)[3].
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 [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'appuie sur la vraisemblance pénalisée de la structure à la manière du critère BIC [2].

$$P(S|X) \propto P(X|S)P(S) \tag{22}$$

$$\ln(P(S|X)) = \ln(P(X|S)) + \ln(P(S)) + cste \tag{23}$$

$$BIC^* = BIC + \ln(P(S)) \tag{24}$$

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

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

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 $j \sim \mathcal{U}(\{1, \ldots, p\})$:

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

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

- Si $i \notin I_i^j$ (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) :

$$P(S|X) \approx exp(BIC^*(S)) \tag{27}$$

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

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 meilleur 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'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 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 ($\rho = 0.911$).

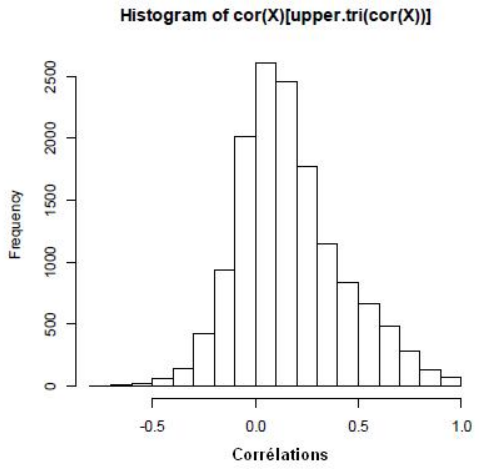
Certaines des sous-régressions obtenues par CorReg ont une interprétation physique:

- 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) :

  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 (modèles ballistiques, etc) utilisés pour réguler le process.

Exemple de régression finale :



	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 $p = 168$. A gauche les corrélations entre variables explicatives, à droite les MSE et la complexité (nombre de variables explicatives) des modèles obtenus pour expliquer Y

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) 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.