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 :

$$
\begin{aligned}
Y_{|X,S} &= XA + \varepsilon_Y = X_1A_1 + X_2A_2 + \varepsilon_Y \text{ avec } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) && (1)\\
\forall j \in I_2 :\ X^j_{|X^{I_1^j},S} &= X^{I_1^j} B^j_{I_1^j} + \varepsilon_j \text{ avec } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) && (2)\\
\forall j \notin I_2 :\ X^j &= f(\theta_j) \text{ mélanges gaussiens orthogonaux à } k_j \text{ composantes} && (3)
\end{aligned}
$$

Où $B^j_{I_1^j}$ est le vecteur de taille p_1^j des coefficients de la sous-régression en X^j,
$I_1 = \{I_1^1, \ldots, I_1^p\}$, $I_2 = \{j | \sharp I_1^j > 0\}$, $X = (X^1, \ldots, X^p) = (X_1, X_2)$ où $X_2 = X^{I_2}$, $A = (A^1, \ldots, A^p) = (A_1, A_2)$ où $A_2 = A^{I_2}$.
On suppose $I_1 \cap I_2 = \emptyset$, *i.e.* les variables dépendantes dans X n'en expliquent pas d'autres.
On note $p_2 = \sharp I_2$ et $p_1 = (p_1^1, \ldots, p_1^{p_2})$.

On a donc rendu explicites les corrélations au sein de X sous la forme d'une structure de sous-régressions linéaires $S = (p_2, I_2, p_1, I_1)$. Cette structure est identifiable au sens de la complexité de la structure et des mélanges gaussiens via un critère de type BIC sous certaines conditions simples. Du point de vue de l'interprétation, les cas non identifiables (triviaux) sont équivalents et ne posent pas de problème.

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

$$
Y_{|X,S} = X_1(A_1 + \sum_{j \in I_2} B^j_{I_1} A_j) + \sum_{j \in I_2} \varepsilon_j A_j + \varepsilon_Y = X_1\alpha_1 + \varepsilon_\alpha = X\alpha + \varepsilon_\alpha \qquad (4)
$$

2 Estimateur

CorReg réduit la variance de l'estimateur en estimant Y seulement à partir de X^{I_1}, sachant (2) et (4). On a ainsi :

$$
\hat{\alpha}_1 = (X_1'X_1)^{-1}X_1'Y \text{ et } \hat{\alpha}_2 = 0 \qquad (5)
$$

estimateur sans biais [3] avec :

$$
\begin{aligned}
\text{Var}\left[\hat{\alpha}_1 | X, S\right] &= (\sigma_Y^2 + \sum_{j \in I_2} \sigma_j^2 A_j^2)(X_1'X_1)^{-1} && (6)\\
& && (7)
\end{aligned}
$$

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). La décorrélation se fait au prix d'un bruit blanc supplémentaire $\sum_{j \in I_2} \varepsilon_j A_j$ qui est d'autant plus faible que les corrélations sont fortes.
Ce nouveau modèle 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 : ceux de corrélation, issus de la structure et qui sont à interpréter comme des 0 de redondance d'information (qui ne signifient donc en rien l'indépendance avec Y) et les 0 de sélection, issus de l'éventuelle méthode de sélection de variables (type LASSO) et qui sont à interpréter comme l'indépendance entre la variable explicative concernée et la variable réponse.
Le modèle obtenu est donc sans biais de prédiction pour Y, 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], mais en prenant comme loi a priori sur S une loi uniforme hiérarchique $P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2)$ plutôt qu'une loi uniforme simple.

$$P(S|X) \propto P(X|S)P(S) \tag{8}$$
$$BIC^* = BIC + \ln(P(S)) \tag{9}$$

L'équiprobabilité ainsi supposée des p_2 et p_1^j 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

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 (le package CorReg permet à l'utilisateur de choisir parmi plusieurs types de voisinage).

On fait l'approximation suivante :

$$P(S|X) \approx \exp(BIC^*(S)) \tag{10}$$

On définit alors

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) = \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_S\}} \frac{\exp(\frac{-1}{2} BIC(\tilde{S}))}{\sum_{S_l \in \mathcal{V}_S} \exp(\frac{-1}{2} BIC(S_l))} \tag{11}$$

$$\tag{12}$$

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

n	p_2	$\sum_{j=1}^{p} \sharp(I_1^j \cap \hat{I_1^j})$	$\sum_{j=1}^{p} \sharp(\hat{I_1^j} \setminus I_1^j)$	$\sum_{j=1}^{p} \sharp(I_1^j \setminus \hat{I_1^j})$	$p_2 - \hat{p_2}$	$\sharp(I_2 \cap \hat{I}_2)$	$\sharp(\hat{I}_2 \setminus I_2)$
30	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
30	8	7.1 (3)	15.13 (7.8)	16.74 (3.1)	3.44 (1)	2.64 (1.2)	1.92 (1.1)
30	16	11.41 (4.2)	28.39 (11.1)	36.25 (4.2)	8.77 (1.2)	4.72 (1.5)	2.51 (1.4)
30	24	17.49 (5)	42.69 (16.1)	54.07 (5.1)	14.03 (1.3)	7.78 (2)	2.19 (1.5)
30	32	19.89 (5.6)	69.28 (14.8)	75.5 (5.6)	18.48 (1.3)	12.35 (1.4)	1.17 (1)
50	0						
50	8						
50	16						
50	24						
50	32						
400	32						

Table 1: Comparaison de la vraie structure et de celle proposée par CorReg. Les variables dépendentes sont bien choisies mais les sous-régressions sont bruitées. Ici, les p_1^j non nuls valent tous 3

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

En pratique, on commence par estimer pour chaque variable de X sa densité sous l'hypothèse d'un mélange gaussien. On peut alors calculer la loi jointe de X pour chaque structure réalisable rencontrée durant l'algorithme MCMC.

4 Résultats

Pour chaque simulation présentée ci-dessous, chacune des configurations à été simulée 100 fois. Les tableaux affichent les moyennes observées et écarts-types. Pour l'ensemble des simulations $p = 40$, les X indépendants suivent des mélanges gaussiens à $\lambda = 5$ classes de moyenne selon une loi de poisson de paramètre λ et d'écart-type λ. Les $B_{i,j}$ suivent la même loi de poisson mais avec un signe aléatoire. On cherche ici à se comparer à la méthode *LASSO* dans les cas où celle-ci est en difficulté (corrélations 2 à 2) donc les p_1^j non nuls valent tous 1. CorReg a travaillé avec K et p_1 libres. Y dépend de 15 variables choisies uniformément dans X sans tenir compte de la structure.

Le tableau 2 montre que CorReg est égal au LASSO en l'absence de structure et le bat quand les corrélations sont fortes. On note également que les modèles proposés par CorReg sont parfois plus simples que le vrai modèle (15 variables) car celui-ci possède une forme réduite (estimée par CorReg). ON retrouve également le phénomène attendu du LASSO moins impacté par les corrélations quand n grandit. On constate enfin la convergence asymptotique de CorReg vers le vrai modèle, comme pour le LASSO.

Les données industrielles sont fortement corrélées de manière naturelle : largeur et

n	p_2	♯ LARS	♯ CorReg	LARS	CorReg	τ
30	0	25.49 (3)	25.49 (3)	**472 385 (1 327 228)**	**472 385 (1 327 228)**	**1**
30	8	20.88 (6.3)	19.4 (6.8)	**519.18 (401.4)**	864.18 (2093.9)	**0.57**
30	16	23.83 (3.8)	18.81 (3.6)	25 671.62 (191 415.1)	**18 843.54 (184 681.9)**	**0.82**
30	24	22.34 (4.3)	16.91 (2.3)	1 163.62 (3 944.4)	**920.79 (4 206.2)**	**0.8**
30	32	17.83 (6.4)	12.1 (2.5)	779.32 (1494.9)	**201.41 (95.5)**	**0.81**
50	0	24.53 (4.7)	24.53 (4.7)	**247.05 (121.9)**	**247.05 (121.9)**	**1**
50	8	16.31 (6.1)	13.71 (3.9)	**183.38 (73.1)**	1808.79 (14 178.1)	**0.66**
50	16	20.11 (4)	16.71 (2.8)	**201.11 (81.3)**	268.63 (1021.1)	**0.77**
50	24	20.61 (4)	16.57 (2.5)	**199.36 (99.6)**	242.14 (842.3)	**0.83**
50	32	15.58 (6)	11.46 (1.9)	185.48 (127.3)	**137.85 (25.6)**	**0.76**
400	32	14.01 (3.6)	11.22 (1.8)	104.96 (4.8)	**103.51 (4.7)**	**0.76**

Table 2: Efficacité en prédiction (echantillons de validation de 1 000 individus) de CorReg par rapport au LASSO suivant le LARS. τ est le taux de victoires (ou égalité) de CorReg, $\sigma_j = 0.01$ et $\sigma_Y = 10$.

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) :

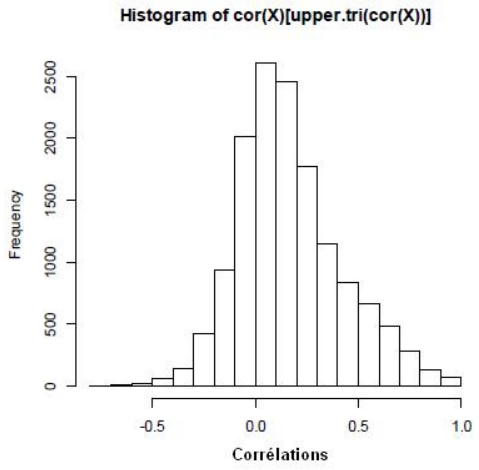
  Largeur = $\frac{\text{débit}}{\text{vitesse} \times \text{épaisseur}}$ (Mais dans ce cas précis l'épaisseur est constante)

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 :

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.



	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 $2,5$.

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.

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.