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 décorrélées, hypothèse souvent irréaliste pour les bases de données d'origine industrielle où de nombreuses corrélations sont dues au process, à des lois physiques, *etc.* Le modèle proposé explicite les corrélations présentes sous la forme d'une famille de régressions linéaires entre covariables, permettant d'obtenir par marginalisation un modèle de régression 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 qui sera illustrée sur données simulées et sur données réelles issues de l'industrie sidérurgique.

Mots-clés. Régression, corrélations, industrie, sélection de variables, modèles génératifs, ...

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, ...

1 Introduction

La régression linéaire classique suppose la décorrélation des covariables, source de problèmes en termes de variance des estimateurs.En effet, pour une variable réponse$Y$ et un ensemble de covariables $X \in \mathcal{R}^p$, la régression $Y = XA + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ donne un estimateur de variance $\mathrm{Var}\,(\hat{A}|X) = \sigma^2(X'X)^{-1}$ dégénéré si les colonnes de X sont linéairement

corrélées. Les méthodes de sélection comme le LASSO [4] muni du LAR [1] sont elles-mêmes touchées par ce problème de corrélation [5].

Notre idée est de modéliser explicitement les corrélations présentes entre covariables sous la forme d'une famille de régression entre celles-ci. Nous présenterons donc le modèle génératif associé et l'algorithme MCMC permettant de choisir la famille de régression à utiliser avant d'illustrer l'efficacité de la méthode sur des données simulées puis sur des données réelles.

2 Modèle supprimant les co-variables corrélées

On suppose le modèle génératif suivant :

- Régression principale :

$$Y_{|X,S} = XA + \varepsilon_Y = X_1A_1 + X_2A_2 + \varepsilon_Y \text{ avec } \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2) \tag{1}$$

- Famille de p_2 régressions entre covariables corrélées :

$$\forall j \in I_2 : \; X^j_{|X_1,S} = X_1B_1^j + \varepsilon_j \text{ avec } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) \tag{2}$$

- Mélanges gaussiens indépendants pour les covariables non corrélées :

$$\forall j \notin I_2 : \; X^j \sim \sum_{k=1}^{k_j} \pi_k \mathcal{N}(\mu_{k_j}, \sigma_{k_j}^2) \tag{3}$$

Où $I_1 = \{I_1^1, \ldots, I_1^p\}$ est le vecteur des indices des variables à droite dans (2) , $I_2 = \{j | \sharp I_1^j > 0\}$ est l'ensemble des indices des variables corrélées à gauche dans (2). On a donc une partition des données $X = (X_1, X_2)$ où $X_2 = X^{I_2}$ et $X_1 = X \setminus X_2$(pour alléger les notations).
On suppose $I_1 \cap I_2 = \emptyset$, *i.e.* les variables dépendantes dans X n'en expliquent pas d'autres. On note $p_2 = \sharp I_2$ le nombre de régressions entre covariables et $p_1 = (p_1^1, \ldots, p_1^{p_2})$.

On a ainsi rendu explicites les corrélations au sein de X sous la forme d'une structure de sous-régressions linéaires $S = (I_1, I_2, p_1, p_2)$. Ce modèle génératif est identifiable sous certaines conditions simples (sur les k_i) non détaillées ici.

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

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

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

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

estimateur sans biais [4] avec :

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

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 [3], 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.

$$P(S|X) \;\propto\; P(X|S)P(S) \tag{7}$$
$$BIC^* \;=\; BIC + \ln(P(S)) \tag{8}$$

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 :

$$P(S|X) \approx \exp(BIC^*(S)) \tag{9}$$

On définit alors

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : \mathcal{P}(S, \tilde{S}) = \frac{1}{p} \sum_{j=1}^{p} \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_S\}} \frac{\exp(\frac{-1}{2} BIC(\tilde{S}))}{\sum_{S_l \in \mathcal{V}_S} \exp(\frac{-1}{2} BIC(S_l))} \tag{10}$$

$$\tag{11}$$

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 [2]. 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).

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 les simulations présentées dans les tableaux 1, 2 et 3, chacune des configurations à été simulée 100 fois. Les tableaux affichent le nombre de variable dépendantes trouvées, le nombre de variables jugées dépendantes à tort et les erreures moyennes en prédiction (MSE) sur Y à partir d'échantillons de validation de 1 000 individus. Pour l'ensemble des simulations $p = 40$, $\sigma_Y = 10$, $\sigma = 0.001$, 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 dans le vrai modèle. CorReg a travaillé avec K,p_2 et p_1 libres.

Les résultats (tableaux 1 à 3) montrent que CorReg est équivalent au LASSO en l'absence de corrélations et le bat quand les corrélations sont fortes. On retrouve 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.

n	p_2	bon gauche	faux gauche	LAR	CorReg $\hat{S}$	CorReg vrai S
30	16	8	5.39	2 466 225.35	13 796.03	588.9
30	32	17.05	2.7	979.16	196.33	141.2
50	0	0	0	499.18	499.18	499.18
50	16	9.18	4.94	315.34	202.64	193.38
50	32	19.13	2.24	179.89	138.96	120.21
400	32	23.66	1.13	105.38	103.88	102.81

Table 1: Y ne dépend pas de X_2.CorReg gagne même en se trompant un peu sur S.

n	p_2	bon gauche	faux gauche	LAR	CorReg $\hat{S}$	CorReg vrai S
30	16	8.48	4.88	3 511 185.23	10 686.62	738.89
30	32	16.89	2.78	565.51	189.54	139.24
50	0	0	0	529.94	529.94	529.94
50	16	8.89	5.4	347.59	233.99	197.95
50	32	18.95	2.44	163.7	139.39	121.56
400	32	23.49	1.06	104.52	103.6	102.67

Table 2: Y dépend de X entier.

n	p_2	bon gauche	faux gauche	LAR	CorReg $\hat{S}$	CorReg vrai S
30	16	8.29	5	5 851.45	559.58	340.29
30	32	17	2.59	893	196.01	135.78
50	16	8.98	5.19	201.56	164.58	162.49
50	32	19.05	2.32	172.93	136.77	121.19
400	32	23.51	1.09	104.49	103.02	102.26

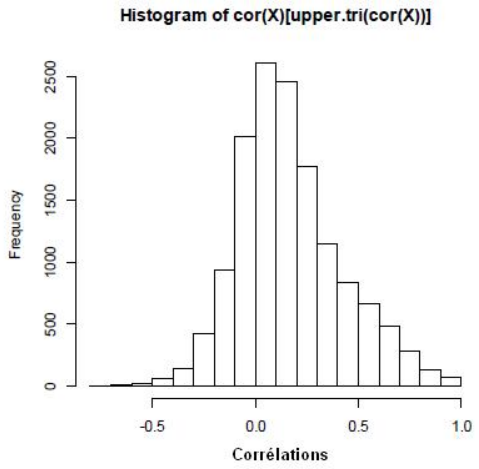
Table 3: Y dépend de X_2 uniquement (cas normalement défavorable à CorReg).

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 industriel sur la base d'expertises métallurgiques.

Exemple de régression sur une variable réponse dans le cadre de 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 sur R-forge. 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 de régression simples et donc 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. En effet, le modèle génératif actuel permettrait cette nouvelle fonctionnalité sans hypothèse supplémentaire.

Bibliographie

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of statistics*, 32(2):407-499.
[2] J. Friedman, T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441 .
[3] E. Lebarbier and T. Mary-Huard (2006). Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS* , 147(1):39-57.
[5] R. Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B (Methodological), pages 267-288.
[6] Peng Zhao and Bin Yu (December 2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* 7:2541-2563.