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 estimée 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 to have decorrelated covariates. This hypothesis is often irrealist with industrial datasets that 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 family of linear regressions between the covariates permitting to obtain by marginalization a parsimonious correlation-free regression model, easily understandable and compatible with variable selection methods. The structure of correlations is found with an MCMC algorithm. An R package (CorReg) implements this new method which will be illustrated on both simulated datasets and real-life datasets from steel industry.

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 $Y \in \mathcal{R}^n$ et un ensemble de covariables $X \in \mathcal{R}^{n \times p}$, la régression $Y = XA + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ (où I_n est la matrice identité de taille n) et $A \in \mathcal{R}^p$ vecteur des p coefficients donne un

estimateur $\hat{A}$ de variance $\mathrm{Var}\,(\hat{A}|X) = \sigma_Y^2(X'X)^{-1}$ 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 régressions entre celles-ci. Nous présenterons donc le modèle génératif associé puis en partie 3 l'algorithme MCMC permettant d'estimer la famille de régressions à utiliser avant d'illustrer dans les parties 4 et 5 l'efficacité de la méthode sur des données simulées puis sur des données réelles avant de conclure en partie 6.

2 Modèle supprimant les covariables corrélées

On suppose le modèle génératif suivant :

- Régression principale entre Y et X:

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

- Famille de p_2 régressions entre covariables de X corrélées :

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

- Mélanges gaussiens indépendants pour les covariables non corrélées :

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

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). Les $B_1^j \in \mathcal{R}^{(p-p_2)}$ sont les coefficients des régressions entre covariables. On a donc une partition des données $X = (X_1, X_2)$ où $X_2 = X^{I_2}$ et $X_1 = X \setminus X_2$. On suppose en outre $I_1 \cap I_2 = \emptyset$, *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 $p_1 = (p_1^1, \ldots, p_1^{p_2})$ qui est le vecteur des longueurs des régressions au sein de X.

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_j) non détaillées ici.

On remarque alors que (1) et (2) impliquent par simple intégration sur X_2, un modèle de régression en Y s'exprimant *uniquement en fonction des variables non corrélées* X_1 :

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

L'estimateur classique du Maximum de Vraisemblance de α est sans biais et s'exprime par

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

En particulier sa matrice de variance

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

peut être notablement mieux conditionnée que celle de $\hat{A}$ initial (dimension réduite et surtout variables orthogonales). En outre, ce nouveau modèle réduit consiste en une régression linéaire classique qui peut donc bénéficier des outils de sélection de variables au même titre que le modèle complet. Notons enfin que la structure explicite entre les variables permet de mieux comprendre les phénomènes en jeu et la parcimonie du modèle facilite son interprétation.

Remarque : En ajoutant une étape de sélection de variables (de type LASSO) on obtient ainsi deux "types de 0" : ceux issus de l'étape de décorrélation et ceux issus de la sélection.

3 Estimation de la structure de corrélation

Le choix de structure s'appuie sur un critère noté BIC^* et qui correspond à la 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. On a donc :

$$BIC^* \quad = \quad BIC + \ln(P(S)). \tag{7}$$

L'équiprobabilité ainsi supposée des p_2 et p_1^j vient pénaliser davantage la compléxité sous l'hypothèse $p_2 < \frac{p}{2}$, hypothèse réaliste sur le nombre de régressions entre covariables. La recherche du meilleur S selon BIC^* n'est pas un problème simple et on va s'appuyer sur un algorithme MCMC pour le résoudre.

A chaque étape de l'algorithme, pour $S \in \mathcal{S}$ (ensemble des structures réalisables) on définit un voisinage $\mathcal{V}_S$ et ensuite la fonction de transition est guidée par BIC^* de la façon suivante :

$$\forall (S, \tilde{S}) \in \mathcal{S}^2 : P(S, \tilde{S}) \quad = \quad \mathbf{1}_{\{\tilde{S} \in \mathcal{V}_S\}} \frac{\exp(-\frac{1}{2} BIC^*(\tilde{S}))}{\sum_{S_l \in \mathcal{V}_S} \exp(-\frac{1}{2} BIC^*(S_l))}. \tag{8}$$

La chaîne de Markov ainsi constituée est ergodique dans un espace d'états finis et possède une unique loi stationnaire dont le mode correspond à la structure de plus grande valeur de BIC^*.

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

		Qualité de $\hat{S}$		Qualité de prédiction (MSE)		
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 1: Y dépend de X entier. CorReg gagne logiquement.

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

En pratique, on commence par estimer pour chaque variable de X sa densité sous l'hypothèse d'un mélange gaussien (avec le package Rmixmod de Mixmod [6]). On peut alors calculer la loi jointe de X pour chaque structure réalisable rencontrée durant l'algorithme MCMC. Sans cette hypothèse générative supplémentaire sur X_1, l'utilisation de BIC^* serait compromise. Notons cependant la souplesse de cette hypothèse due à la grande flexibilité des mélanges gaussiens [7].

4 Résultats sur données simulées

L'ensemble de la méthode à été programmé dans un package R dénommé CorReg. Pour les simulations présentées dans les tableaux 1 et 2, chacune des configurations à été simulée 100 fois. Les tableaux affichent le nombre de variables dépendantes trouvées ("bon gauche"), le nombre de variables jugées dépendantes à tort ("faux gauche") et les erreurs 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_1^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 (fortes corrélations 2 à 2) donc les p_1^j non nuls valent tous 1 dans le vrai modèle. CorReg a travaillé avec p_2 et p_1 libres et a utilisé Mixmod pour estimer les densités dans X_1.

Les résultats (tableaux 1 et 2) 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 de régression.

On remarque que quand p_2 augmente le LASSO commence à se ressaisir car il y a de plus en plus de faux modèles proches du vrai en termes de prédiction. Le LASSO trouve

		Qualité de $\hat{S}$		Qualité de prédiction (MSE)		
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 2: Y dépend de X_2 uniquement (cas normalement défavorable à CorReg).

donc par moment des modèles inconsistants en interprétation mais relativement corrects en prédiction.

5 Résultats sur données réelles

La figure 1 illustre les résultats obtenus lors d'une étude qualité chez ArcelorMittal. Les données sidérurgiques 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 ($\rho = 0.911$). La base analysée ici comportait 205 variables et 3 847 individus (apprentissage du modèle sur les 3 000 premiers et validation sur les 847 autres). CorReg retrouve ($p_2 = 82$) en plus des corrélations ci-dessus des modèles de régulation du process et certains modèles physiques naturels.
Le MSE (sur l'échantillon de validation) obtenu par CorReg est **6.47**% meilleur que celui du LASSO. CorReg donne un modèle à 34 variables contre 21 pour le LASSO avec 14 variables communes. L'étude a été menée dans le cadre de la qualité produit. L'enjeu de ces quelques pourcents de gain se chiffre en dizaine de milliers d'euros annuels sans compter l'impact sur les parts de marché (non chiffrable mais bien plus considérable).

6 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

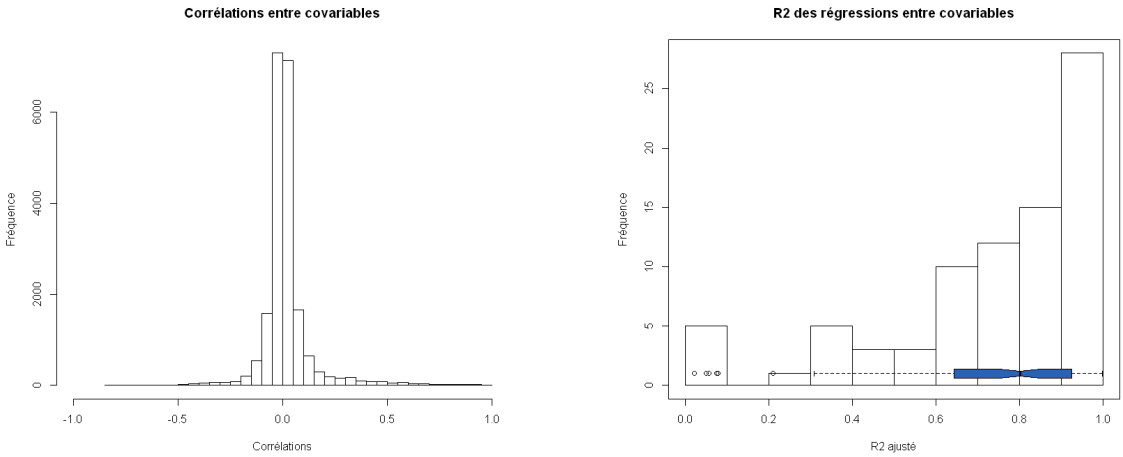

Figure 1: Histogrammes des corrélations et des R^2_{adj} des régressions obtenues par CorReg

actuel permettrait cette nouvelle fonctionnalité sans hypothèse supplémentaire. Enfin, la famille de régressions pourrait être utilisée dans domaines autres que celui de la régression linéaire. La connaissance de la structure interne des données possède en effet une valeur intrinsèque que d'autres méthodes pourraient tenter de mettre à profit.

Bibliographie

[1] Efron, B., Hastie, T., Johnstone,I. et Tibshirani, R. (2004), Least angle regression. *The Annals of statistics*, 32(2):407-499.

[2] Friedman, J., Hastie, T. et Tibshirani, R. (2008), Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432-441 .

[3] Lebarbier, E. et Mary-Huard,T. (2006), Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la SFdS* , 147(1):39-57.

[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society.* Series B (Methodological), pages 267-288.

[5] Zhao,P. et Yu,B. (2006), On model selection consistency of lasso, *J. Mach. Learn. Res.* 7:2541-2563.

[6] Biernacki, C., Celeux, G., Govaert, G., & Langrognet, F. (2006), Model-based cluster and discriminant analysis with the MIXMOD software, Computational Statistics & Data Analysis, 51(2), 587-600.

[7] McLachlan, G. J., et Basford, K. E. (1988), Mixture models : Inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker* , 1.