CorReg : Préselection de variables en régression linéaire avec fortes corrélations

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

[1] *ArcelorMittal, Université Lille 1, Inria, CNRS, clement.thery@arcelormittal.com*
[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 est pénalisée par l'usage de variables explicatives corrélées, situation fréquente pour les bases de données d'origine industrielle où les corrélations sont nombreuses et mènent à des estimateurs de forte variance. 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 consistant en une préselection de variables. La structure de corrélations est estimée à l'aide d'un algorithme MCMC qui repose sur un modèle génératif complet. Le package CorReg (sur le CRAN) permet la mise en oeuvre en R 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 is penalized by correlated covariates, frequent situation for industrial datasets, in terms variance of the estimators. 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 that can be seen as a variable preselection. The structure of correlations is found with an MCMC algorithm. An R package (CorReg) available on the CRAN 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

Les corrélations entre variables en régression linéaire sont sources de problèmes en termes de variance des estimateurs et de sélection de variables. En effet, pour une variable réponse $\boldsymbol{Y} \in \mathcal{R}^n$ et un ensemble de covariables $\boldsymbol{X} \in \mathcal{R}^{n \times p}$, la régression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ (où $\boldsymbol{I}_n$ est la matrice identité de taille n) et $\boldsymbol{\beta} \in \mathcal{R}^p$ vecteur des p coefficients donne un estimateur $\hat{\boldsymbol{\beta}}$ de variance $\mathrm{Var}\,(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_Y^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ dégénéré si les

colonnes de $\boldsymbol{X}$ 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. L'estimation de cette famille consiste en un choix de modèle génératif pour les variables explicatives à l'aide d'un algorithme MCMC que nous présentons en partie 3 avant d'illustrer dans les parties 4 et 5 l'efficacité de la méthode sur données simulées puis sur données réelles avant de conclure.

2 Modèle supprimant les covariables corrélées

On suppose le modèle génératif suivant :

- Régression principale entre $\boldsymbol{Y}$ et $\boldsymbol{X}$:

$$\boldsymbol{Y}_{|\boldsymbol{X},S} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_Y = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_Y \text{ avec } \boldsymbol{\varepsilon}_Y \sim \mathcal{N}(\boldsymbol{0}, \sigma_Y^2 \boldsymbol{I}_n); \tag{1}$$

  On rend alors explicites les corrélations au sein de $\boldsymbol{X}$ sous la forme d'une structure de sous-régressions linéaires $S = (I_1, I_2, p_1, p_2)$.

- Famille de p_2 régressions entre covariables de $\boldsymbol{X}$ corrélées :

$$\forall j \in I_2 : \ \boldsymbol{X}^j_{|\boldsymbol{X}_1,S} = \boldsymbol{X}_1\boldsymbol{\alpha}^j + \boldsymbol{\varepsilon}_j \text{ avec } \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\boldsymbol{0}, \sigma_j^2 \boldsymbol{I}_n); \tag{2}$$

où I_2 est l'ensemble des indices des variables corrélées à gauche dans (2) et $I_1 = \{I_1^1, \ldots, I_1^p\}$ est l'ensemble des ensembles des indices des variables à droite dans (2), avec $I_1^j = \emptyset$ si $j \notin I_2$. Les $\boldsymbol{\alpha}^j \in \mathcal{R}^{(p-p_2)}$ sont les coefficients des régressions entre covariables. On a donc une partition des données $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ où $\boldsymbol{X}_2 = \boldsymbol{X}^{I_2}$ et $\boldsymbol{X}_1 = \boldsymbol{X} \setminus \boldsymbol{X}_2$. On suppose en outre $I_1 \cap I_2 = \emptyset$, *i.e.* les variables dépendantes dans $\boldsymbol{X}$ n'en expliquent pas d'autres. On note $p_2 = \sharp I_2$ le nombre de régressions entre covariables et $p_1 = (p_1^1, \ldots, p_1^p)$ qui est le vecteur des longueurs des régressions au sein de $\boldsymbol{X}$ avec $p_1^j = \sharp I_1^j$.

On remarque alors que (1) et (2) impliquent par simple intégration sur $\boldsymbol{X}_2$, un modèle marginal de régression en $\boldsymbol{Y}$ s'exprimant *uniquement en fonction des variables non corrélées* $\boldsymbol{X}_1$:

$$\boldsymbol{Y}_{|\boldsymbol{X}_1,S} \ = \ \boldsymbol{X}_1(\boldsymbol{\beta}_1 + \sum_{j \in I_2} \boldsymbol{\beta}_j \boldsymbol{\alpha}_j) + \sum_{j \in I_2} \boldsymbol{\varepsilon}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_Y = \boldsymbol{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}_Y^*. \tag{3}$$

En outre, ce nouveau modèle marginal consiste en une régression linéaire classique qui peut donc bénéficier des outils de sélection de variables au même titre que le modèle complet. On a ainsi un prétraitement sur les données par préselection visant à décorréler les

variables utilisées dans le modèle en $\boldsymbol{Y}$. L'estimateur classique du Maximum de Vraisemblance de $\boldsymbol{\beta}^*$ est sans biais et s'écrit

$$\hat{\boldsymbol{\beta}}_1^* = (\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1}\boldsymbol{X}'_1\boldsymbol{Y} \tag{4}$$

En particulier sa matrice de variance

$$\text{Var}\,[\hat{\boldsymbol{\beta}}_1^*|\boldsymbol{X}, S] = (\sigma_Y^2 + \sum_{j\in I_2}\sigma_j^2\boldsymbol{\beta}j^2)(\boldsymbol{X}'_1\boldsymbol{X}_1)^{-1} \tag{5}$$

peut être notablement mieux conditionnée que celle de $\hat{\boldsymbol{\beta}}$ initial (dimension réduite et surtout variables orthogonales).

Enfin, la structure explicite 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

Pour choisir parmi des structures de taille différente, on s'appuye sur $P(S|\boldsymbol{X})$ qui est proportionnel à $P(\boldsymbol{X}_2|\boldsymbol{X}_1, S)P(\boldsymbol{X}_1|S)P(S)$. On fait alors l'hypothèse de mélanges gaussiens indépendants pour les covariables non corrélées :

$$P(\boldsymbol{X}_1|S) : \forall j \notin I_2 : \ \boldsymbol{X}^j \sim \sum_{k=1}^{k_j}\boldsymbol{\pi}_k\mathcal{N}(\boldsymbol{\mu}_{k_j}, \boldsymbol{\sigma}_{k_j}^2\boldsymbol{I}_n); \tag{6}$$

On prend ensuite comme loi *a priori* sur S, plutôt qu'une loi uniforme simple, une loi uniforme hiérarchique $P(S) = P(I_1|p_1, I_2, p_2)P(p_1|I_2, p_2)P(I_2|p_2)P(p_2)$. L'équiprobabilité ainsi supposée des p_2 et p_1^j vient pénaliser la compléxité sous l'hypothèse $p_2 < \frac{p}{2}$, hypothèse réaliste sur le nombre de régressions entre covariables. La recherche du meilleur S est combinatoire et un algorithme MCMC est utilisé par souci d'efficacité et de flexibilité. On optimise alors un critère de type BIC [3], noté BIC* :

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

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^* selon :

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

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 qui optimise BIC^*.

		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: $\boldsymbol{Y}$ dépend de $\boldsymbol{X}$ entier. CorReg gagne logiquement.

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 [8] (pour 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 $\boldsymbol{X}$ sa densité sous l'hypothèse d'un mélange gaussien (package Rmixmod de Mixmod [6]). On peut alors ensuite calculer la loi jointe de $\boldsymbol{X}$ pour chaque structure réalisable rencontrée durant l'algorithme MCMC. 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 pour R (package 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 $\boldsymbol{Y}$ à partir d'échantillons de validation de 1 000 individus. Pour l'ensemble des simulations $p = 40$, $\sigma_Y = 10$, $\sigma = 0.001$, les $\boldsymbol{X}$ 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 $\boldsymbol{\alpha}_j$ 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 le vrai modèle est constitué de corrélations 2 à 2. CorReg a travaillé avec p_2 et p_1 libres.

Les 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 donc le LASSO trouve des modèles inconsistants en interprétation mais relativement corrects en prédiction.

		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: $\boldsymbol{Y}$ dépend de $\boldsymbol{X}_2$ uniquement (cas normalement défavorable à CorReg).

5 Résultats d'une étude qualité chez ArcelorMittal

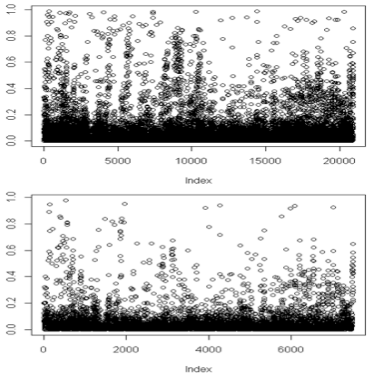

Figure 1: Valeurs de ρ pour $\boldsymbol{X}$ (haut) et $\boldsymbol{X}_1$ (bas).

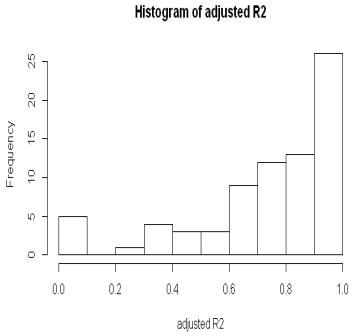

Figure 2: R^2_{adj} des 82 régressions obtenues.

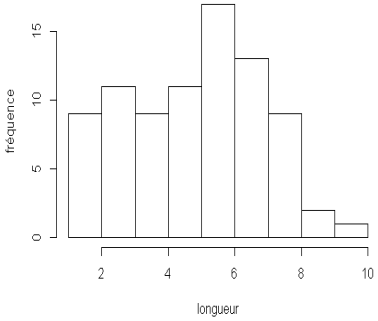

Figure 3: Longueur des régressions obtenues (p_1).

On note ρ la valeur absolue des corrélations. Les données sidérurgiques étudiées (p=205 et n=3000) sont fortement corrélées de manière naturelle (Figure 1 en haut), comme la largeur et poids d'une brame (ρ=0.905), la température avant et après un outil ($\rho = 0.983$), la rugosité des deux faces du produit (ρ=0.919), une moyenne et un maximum (ρ=0.911). CorReg trouve en plus des corrélations ci-dessus des modèles de régulation du process et des modèles physiques naturels pour un total de $p_2 = 82$ régressions (Figure 2) de longueur moyenne $\bar{p}_1 = 5$ (Figure 3). Entre $\boldsymbol{X}$ et $\boldsymbol{X}_1$ le nombre de $\rho > 0.7$ est réduit de $\mathbf{79,33}\%$ avec respectivement 150 et 31 paires de variables (Figure 1 en bas).
Ici $\boldsymbol{Y}$ est un indicateur qualité produit (confidentiel). Le MSE (sur échantillon de validation de 847 nouveaux individus) obtenu par CorReg est $\mathbf{1.55}\%$ meilleur que celui du LASSO, avec respectivement 31 et 20 variables dont 13 communes. LASSO propose 7 variables différentes de CorReg mais elles sont toutes dans $\boldsymbol{X}_2$ et CorReg reprend les variables explicatives des régressions correspondantes (R^2_{adj} moyen de 0.82). De plus ρ est

13.9% plus faible pour les variables de CorReg malgré davantage de variables.
En termes d'interprétation, accompagner la régression en $\boldsymbol{Y}$ avec la famille de régressions permet de mieux comprendre les conséquences d'éventuelles mesures correctives sur l'ensemble du process. Cela permet typiquement de déterminer les *actionneurs* qui influent sur $\boldsymbol{Y}$ quand le LASSO fait ressortir des variables *subies*. On peut donc plus facilement corriger le process pour atteindre l'objectif fixé. 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 disponible sur le CRAN et a d'ores et déjà montré son efficacité sur de vraies problématiques de régression en entreprise. Sa force est la grande interprétabilité du modèle proposé, qui est constitué de plusieurs régression linéaires courtes et donc facilement accessibles aux non statisticiens 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, aussi très présentes dans l'industrie. D'ailleurs le modèle génératif actuel permettrait cette nouvelle fonctionnalité sans hypothèse supplémentaire, ce qui renforce encore son intérêt. Enfin, le principe de CorReg qui est l'explicitation des régressions latentes entre covariables pourrait être appliqué à d'autres méthodes prédictives (régression logistique, *etc.*).

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., et Langrognet, F. (2006), Model-based cluster and discriminant analysis with the MIXMOD software, Computational Statistics & Data Analysis, 51(2), 587-600.
[7] McLachlan, G., et Peel, D. (2004). Finite mixture models. Wiley. com.
[8] Gilks, W. R., Richardson, S., et Spiegelhalter, D. J. (Eds.). (1996). Markov chain Monte Carlo in practice (Vol. 2). CRC press.