Regression for correlated variables : Application in steel industry

Clément Théry [1]

[1] *ArcelorMittal Dunkerque, Inria Lille, Universit 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 expliciter les corrélations présentes sous la forme d'une 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 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 rgression linaire classique suppose l'indpendance des covariables. Les corrlations sont problmatiques et posent des problmes.

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

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

2 Le modle gnratif

3 Estimateur

4 Recherche de structure

5 Slection de variables

6 Rsultats

7 Conclusion et perspectives

CorReg est fonctionnel et disponible Besoin d'largir la gestion des valeurs manquantes trs prsentes dans l'industrie

8 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

[1]

References

[1] D.W. Marquardt and R.D. Snee. Ridge regression in practice. *American Statistician*, pages 3–20, 1975.

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.