Régression non linéaire robuste en grande dimension

Emeline Perthame, Florence Forbes, Brice Olivier, Antoine Deleforge

Régression non linéaire robuste en grande dimension

Emeline Perthame [1] & Florence Forbes [1] & Brice Olivier [1] & Antoine Deleforge [2]

[1] *INRIA Grenoble Rhone-Alpes*
655 avenue de l'Europe
38334 Montbonnot Cedex
E-mail : emeline.perthame@inria.fr & florence.forbes@inria.fr & brice.olivier@inria.fr

[4] *INRIA Rennes - Bretagne Atlantique*
Campus universitaire de Beaulieu
35042 Rennes Cedex
antoine.deleforge@inria.fr

Résumé. La régression non-linéaire permet de modéliser des relations complexes entre des variables cibles et un nombre possiblement grand de covariables. Cependant, dans le cadre classique gaussien, il a été montré que les outliers affectent la stabilité des résultats ce qui peut mener à des prédictions erronées. Il est donc nécessaire de développer des approches robustes, applicables en grande dimension, afin de réduire l'impact de ces outliers et d'améliorer la précision des méthodes de régression linéaire ou non linéaire.

La non-linéarité est prise en compte dans la méthode proposée par un modèle de mélange de régressions. Les modèles de mélanges, et paradoxalement les mélanges de régression sont principalement utilisés pour répondre à un objectif de classification et peu d'articles font référence aux mélanges de régression dans une optique de régression et de prédiction. La pertinence d'une approche de prédiction fondée sur un mélange de régression dans un contexte Gaussien a pourtant été montrée dans (Deleforge et al., 2015 [1]). Cependant la méthode développée par ces auteurs n'est pas une approche de régression robuste. On propose donc d'étendre cette méthode en considérant un mélange de lois de Student généralisées, capables de prendre en compte les outliers. Un algorithme EM est proposé pour l'estimation des paramètres, numériquement implémentable en grande dimension (nombre de variables supérieur au nombre d'observations). Durant la présentation, les performances de la méthode seront étudiées sur des simulations et sur une application sur données réelles.

Mots-clés. Régression non linéaire, régression robuste, grande dimension

Abstract. Non linear regression is used to model complex relations between a target and a possibly large number of features. Nevertheless, under the common gaussian setting, outliers are known to affect the stability of the results and can lead to misleading predictions. Robust approaches that are tractable in high dimension are therefore needed

in order to improve the accuracy of linear or non-linear regression methods under the presence of outliers.

In the proposed method, non linearity is handled via a mixture of regressions. Mixture models and paradoxically also the so-called mixture of regression models are mostly used to handle clustering issues and few articles refer to mixture models for actual prediction purposes. Interestingly, it was shown in (Deleforge et al., 2015 [1]) that a prediction approach based on mixture of regressions in a Gaussian setting was relevant. However, the method developed by these authors is not designed to perform robust regression. Therefore, we build on the work in [1] by considering mixture of Student distributions that are able to handle outliers.

The parameter estimation can be performed via an EM algorithm which remains numerically feasible when the number of variables exceeds the number of observations. During the talk, intensive simulations, both on illustrative and more complex examples in high dimension, will demonstrate that the proposed model performs well in this setting. Application of the method on real datasets will also be illustrated.

Keywords. Non linear regression, robust regression, high dimension

1 Introduction

L'analyse de spectres issus d'études protéomiques en biologie ou de mesures infra-rouges en chimiométrie implique souvent des relations non-linéaires entre les variables réponses et les variables explicatives. Dans ce contexte, la régression non-linéaire permet de modéliser les relations complexes entre les variables cibles et un nombre possiblement grand de covariables. A l'instar de la régression linéaire, les observations sont souvent supposées distribuées selon une loi normale. Cependant, sous cette hypothèse, il a été montré que la stabilité des résultats est affectée par la présence d'outliers ce qui peut mener à des prédictions erronées. Il est donc nécessaire de développer des approches robustes, applicables en grande dimension, afin de réduire l'impact des outliers et d'améliorer la précision des méthodes de régression linéaire ou non linéaire.

On considère le problème de régression en grande dimension suivant : pour l'observation $n \in \{1, \ldots, N\}$, $X_n \in \mathbb{R}^L$ désigne un vecteur des variables réponses de dimension L et $Y_n \in \mathbb{R}^D$ désigne un vecteur de variables explicatives de dimension D, avec $L \ll D$. N peut être inférieur à D. L'objectif est d'estimer la fonction de régression g suivante :

$$\mathbb{E}(X_n|Y_n = y_n) = g(y_n). \tag{1}$$

2 Modèle

Compte tenu de la dimension de Y_n de l'ordre de plusieurs centaines, le principe de la méthode repose sur la régression inverse de Y sur X. Cette technique permet de réduire considérablement le nombre de paramètres à estimer et la complexité du modèle.

La non-linéarité de la relation entre X et Y est prise en compte dans la méthode proposée par un modèle de mélange de régressions (voir [2]). Les modèles de mélanges, et paradoxalement les mélanges de régression sont principalement utilisés pour répondre à un objectif de classification et peu d'articles font référence aux mélanges de régression dans une optique de régression et de prédiction. La pertinence d'une approche de prédiction fondée sur un mélange de régressions dans un contexte Gaussien a pourtant été montrée dans (Deleforge et al., 2015 [1]). Cependant la méthode GLLiM, introduite par ces auteurs, n'est pas une approche de régression robuste. On propose donc d'étendre cette méthode en considérant un mélange de lois de Student généralisées (voir [3]), capables de prendre en compte les outliers dans le modèle. Cette distribution est définie par la fonction de densité $\mathcal{S}_M(\cdot, \mu, \Sigma, \alpha, \gamma)$ suivante :

$$\mathcal{S}_M(y, \mu, \Sigma; \alpha, \gamma) = \frac{\Gamma(\alpha + M/2)}{|\Sigma|^{1/2}\Gamma(\alpha)(2\pi\gamma)^{M/2}} \frac{1}{[1 + \delta(y, \mu, \Sigma)/(2\gamma)]^{(\alpha+M/2)}}$$

où $\delta(y, \mu, \Sigma) = (y - \mu)'\Sigma^{-1}(y - \mu)$ est le carré de la distance de Mahalanobis de y. Les paramètres μ et Σ sont des paramètres de tendance centrale et de covariance et les degrés de liberté α et γ règlent l'épaisseur de la queue de distribution.

Modélisation jointe de (X_n, Y_n) Dans ce contexte de mélanges de loi de Student, on définit le modèle SLLiM par une modélisation jointe des variables réponses X_n et des variables explicatives Y_n. Pour tout $n \in \{1, \ldots, N\}$, on suppose :

$$p(X_n = x_n | Z_n = k; \theta, \phi) = \mathcal{S}_L(x_n, c_k, \Gamma_k, \alpha_k, \gamma_k)$$

où $c_k \in \mathbb{R}^L$ et $\Gamma_k \in \mathbb{R}^L \times \mathbb{R}^L$ et on suppose la distribution conditionnelle suivante :

$$p(Y_n = y_n | X_n = x_n, Z_n = k; \theta, \phi) = \mathcal{S}_D(y_n, A_k x_n + b_k, \Sigma_k, \alpha_k^y, \gamma_k^y).$$

Cette distribution définit une régression inverse *low-to-high*. Les degrés de liberté α_k^y et γ_k^y dépendent de la dimension de X_n et de la distance de Mahalanobis de x_n:

$$\alpha_k^y = \alpha_k + L/2,$$
$$\gamma_k^y = \gamma_k + \frac{1}{2}\delta(x_n, c_k, \Gamma_k).$$

Enfin, la variable cachée $(Z_n)_{(1 \leq n \leq N)}$ suit une loi multinomiale telle que $p(Z_n = k, \phi) = \pi_k$.

Expression de la fonction de régression A la manière de [1], la grande dimension des données est contrôlée grâce à une régression inverse. Le modèle joint défini ci-dessus permet une expression analytique de la fonction de régression d'intérêt présenté à l'Equation (1):

$$\mathbb{E}(X_n|y_n,\theta,\phi) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{S}(y_n, c_k^*, \Gamma_k^*, \alpha_k, \gamma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{S}(y_n, c_j^*, \Gamma_j^*, \alpha_j, \gamma_j)} (A_k^* y_n + b_k^*)$$

où les paramètres $(A_k^*, b_k^*, c_k^*, \Gamma_k^*)_{1\leq k\leq K}$ admettent des expressions analytiques en fonction des paramètres d'origine $\theta = (A_k, b_k, \Sigma_k, c_k, \Gamma_k)_{1\leq k\leq K}$ et $\phi = (\pi_k, \alpha_k, \gamma_k)_{1\leq k\leq K}$.

L'estimation des paramètres θ et ϕ est possible par un algorithme de type Expectation-Maximization, qui reste performant dans un contexte de grande dimension, lorsque le nombre de variables dépasse le nombre d'observations.

3 Illustration en petite dimension

Le but de l'exemple suivant est d'illustrer que le modèle SLLiM permet de réduire l'impact négatif des outliers sur un exemple représentable graphiquement ($L = D = 1$), par rapport à la méthode existante GLLiM.

Le groupe responsable des transports publics à Paris (RATP) étudie depuis 2013 la qualité de l'air sur le réseau souterrain. La qualité de l'air et le climat sont mesurés, tout au long de la journée. Les données sont accessibles sur `http://data.ratp.fr/explore/dataset/qualite-de-lair-mesuree-dans-la-station-chatelet`. Le jeu de données étudié contient les mesures de la qualité de l'air dans 3 stations, dont Châtelet sur la ligne 4. Chaque jour, à chaque heure, des marqueurs climatiques et le renouvellement d'air sont mesurés ainsi que la qualité de l'air (particules et oxydes d'azote). On s'intéresse ici à la relation entre le dioxyde d'azote et le monoxyde d'azote pendant le mois de mars 2015. Les données contiennent 720 points et sont représentées Figure 1 (points gris). On remarque que ces deux variables sont liées par une relation non linéaire. On remarque aussi la présence d'outliers.

La Figure 1 représente la fonction de régression estimée par les méthodes GLLiM (en bleu) et SLLiM (en noir) pour $K = 2$ sur les données complètes (Figure 1a) et sur les données auxquelles ont été supprimées 6 outliers identifiés visuellement (Figure 1b). On désigne par outliers les observations pour lesquelles le taux de monoxyde d'azote dépasse $200\mu g/m^3$. Sur ces figures, on remarque que la fonction de régression est influencée par les fortes valeurs de NO. Après suppression de 6 observations influentes, la fonction estimé par GLLiM rejoint visuellement celle de SLLiM. Cette sensibilité aux outliers se répercute

Table 1: Erreur de prédiction pour GLLiM et SLLiM, sur les données complètes et après suppression de 6 outliers

Données	Méthode	Erreur de prédiction
Complètes	GLLiM	0.81
	SLLiM	0.44
Incomplètes	GLLiM	0.53
	SLLiM	0.43

sur l'erreur de prédiction. La Table 1 présente l'erreur de prédiction calculée par validation croisée (erreur quadratique normalisée). Sur les données complètes, on remarque que le modèle construit à partir des lois de Student est plus performant en matière de prédiction que le modèle gaussien. On remarque aussi qu'après suppression de 6 outliers, les performances de GLLiM s'approchent de celles de SLLiM, qui elles, restent similaires à celles observées sur les données complètes (0.44 vs. 0.43).

En pratique, la méthode proposée est numériquement faisable en grande dimension ($D > N$). Durant la présentation, les performances de ce modèle seront illustrées sur une étude par simulations en grande dimension. Une application de ce modèle sur des données réelles sera aussi présentée.

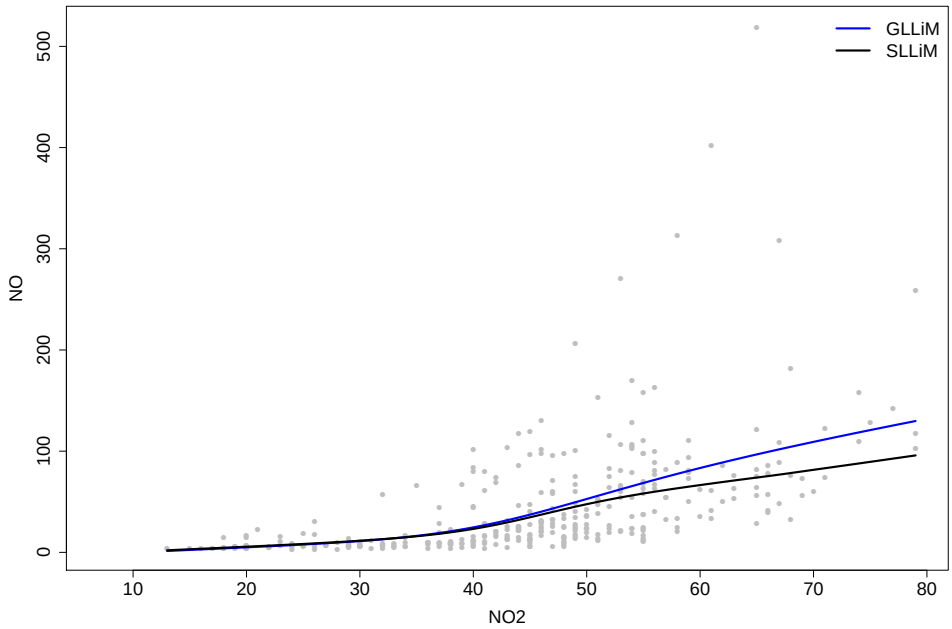
Bibliographie

[1] Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. Statistics and Computing, 25(5):893–911.

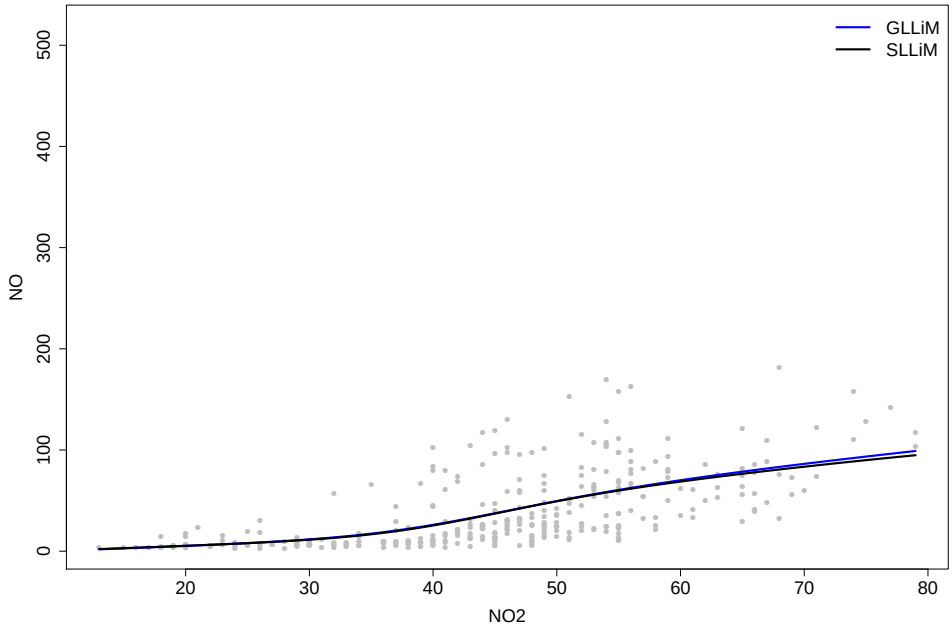
[2] Frühwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. Springer Series in Statistics

[3] Kotz, S. and Nadarajah, S. (2004). Multivariate t Distributions and Their Applications. Cambridge University press

Figure 1: Taux de monoxyde d'azote en fonction du taux de dioxyde d'azote durant le mois de mars 2015 à la station Châtelet - Nuage de points et fonction de régression estimée par mélange gaussien (GLLiM) et de Student (SLLiM)



(a) Prédiction sur les données complètes



(b) Prédiction après suppression de 6 outliers