HAL open science

Régularisation dans les Modèles Linéaires Généralisés Mixtes avec effet aléatoire autorégressif

Jocelyn Chauvet, Catherine Trottier, Xavier Bry

To cite this version:

Jocelyn Chauvet, Catherine Trottier, Xavier Bry. Régularisation dans les Modèles Linéaires Généralisés Mixtes avec effet aléatoire autorégressif. JdS 2017, 49èmes Journées de Statistique de la SFdS, May 2017, Avignon, France. hal-01818544

HAL Id: hal-01818544

https://hal.science/hal-01818544v1

Submitted on 7 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Régularisation dans les Modèles Linéaires Généralisés Mixtes avec effet aléatoire autorégressif

Jocelyn Chauvet[1], Catherine Trottier[1,2] & Xavier Bry[1]

[1] *Institut Montpelliérain Alexander Grothendieck, CNRS, Univ Montpellier, France.*
jocelyn.chauvet@umontpellier.fr, xavier.bry@univ-montp2.fr.
[2] *Univ Paul-Valéry Montpellier 3, Montpellier, France.*
catherine.trottier@univ-montp3.fr.

Résumé. Nous proposons des versions régularisées de l'algorithme Espérance - Maximisation (EM) permettant d'estimer un Modèle Linéaire Généralisé Mixte (GLMM) pour des données de panel (mesurées sur plusieurs individus à différentes dates). Une réponse aléatoire y est modélisée par un GLMM, au moyen d'un ensemble X de variables explicatives et de deux effets aléatoires. Le premier effet modélise la dépendance des mesures relatives à un même individu, tandis que le second représente l'effet temporel autocorrélé partagé par tous les individus. Les variables dans X sont supposées nombreuses et redondantes, si bien qu'il est nécessaire de régulariser la régression. Dans ce contexte, nous proposons d'abord un algorithme EM pénalisé en norme L_2 pour des données de petite dimension, puis une version régularisée de l'algorithme EM, basée sur la construction de composantes supervisées, plutôt destinée à la grande dimension.

Mots-clés. Algorithme EM régularisé, Modèles Linéaires Généralisés Mixtes, Effet aléatoire autocorrélé, Données de panel.

Abstract. We address regularised versions of the Expectation-Maximisation (EM) algorithm for Generalised Linear Mixed Models (GLMM) in the context of panel data (measured on several individuals at different time points). A random response y is modelled by a GLMM, using a set X of explanatory variables and two random effects. The first effect introduces the dependence within individuals on which data is repeatedly collected while the second embodies the serially correlated time-specific effect shared by all the individuals. Variables in X are assumed many and redundant, so that regression demands regularisation. In this context, we first propose a L_2-penalised EM algorithm for low-dimensional data, and then a supervised component-based regularised EM algorithm for the high-dimensional case.

Keywords. Regularised EM algorithm, Generalised Linear Mixed Model, Autoregressive random effect, Panel data analysis.

1 Introduction

L'un des objectifs principaux de l'analyse des données de panel est de prendre en compte la dépendance engendrée par la présence de mesures répétées au cours du temps. Par ailleurs, au vu des facilités actuelles pour collecter de grandes masses de données, les fortes corrélations potentielles au sein des variables explicatives doivent également être considérées. Dans ce but, des régularisations de type ridge et lasso ainsi que des méthodes à composantes ont récemment été mises en avant.

Dans le cadre des Modèles Linéaires Mixtes (LMM), Eliot et al. (2011) proposent d'étendre la régression ridge aux données longitudinales. Afin de maximiser une vraisemblance pénalisée sur la norme L_2 des coefficients, ils suggèrent une variante de l'algorithme EM qui inclut, à chaque itération, la détermination du meilleur coefficient de pénalisation au travers d'une étape de Validation Croisée Généralisée (GCV). Une autre méthode basée sur une vraisemblance pénalisée, cette fois-ci dans une perspective de sélection de variables, est proposé par Schelldorfer et al. (2014). Ils élaborent à cet effet un algorithme de type lasso pour ajuster des Modèles Linéaires Généralisés Mixtes (GLMM) de grande dimension, qui combine approximation de Laplace et algorithme de descente de gradient.

Dans le cadre des GLM, Bry et al. (2013) mettent en œuvre une méthode de type PLS – nommée Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR)) – qui régularise le prédicteur linéaire tout en facilitant son interprétation au moyen de composantes explicatives. Inspirés par l'algorithme de Schall (1991), Chauvet et al. (2016) étendent cette stratégie de régularisation aux données groupées, et plus généralement à l'ensemble de la classe des GLMM.

À notre connaissance, les effets aléatoires apparaissant dans les stratégies précédentes sont supposés distribués selon des lois normales avec des niveaux indépendants. Cependant, pour les données de panel, il est naturel de greffer une structure d'autocorrélation à l'effet aléatoire temporel. Deux objectifs complémentaires émergent alors : étendre d'une part la régression ridge de Eliot et al. (2011) aux GLMM avec effet aléatoire AR(1) ; et présenter d'autre part une nouvelle version de SCGLR adaptée pour les données de panel de grande dimension.

2 Modélisation

Dans cette section, nous rappelons les hypothèses principales concernant les GLMM et nous introduisons les distributions des effets aléatoires. Dans un souci de clarté, nous nous focaliserons sur des données de panel équilibrées avec N individus, chacun d'eux observés en T dates. On note $n = N \times T$ le nombre total d'observations, X la matrice de design des effets fixes (de taille $n \times p$) et U celle des effets aléatoires (de taille $n \times q$). Par ailleurs, Y désigne le vecteur de taille n des réponses aléatoires, β le vecteur de taille p des effets fixes, et ξ le vecteur de taille q des effets aléatoires. Nous observons une réalisation y de Y tandis que ξ n'est pas observé. Nous supposons usuellement que :

(i) les $Y_i \,|\, \xi,\ i \in \{1, \ldots, n\}$ sont indépendants et leurs distributions appartiennent à la famille exponentielle ;

(ii) l'espérance conditionnelle $\mu_i = \mathbb{E}(Y_i \,|\, \xi)$ dépend de β et ξ au travers de la fonction de lien g et du prédicteur linéaire $\eta_i = x_i^T \beta + u_i^T \xi$, vérifiant $\eta_i = g(\mu_i)$.

Dans notre modèle, nous considérons deux effets aléatoires ξ_1 et ξ_2, aux rôles et distributions bien différents :

(i) ξ_1 est l'effet aléatoire spécifique aux individus. En les supposant indépendants, on pose :

$$\xi_1 \sim \mathcal{N}_N \left(0,\, \sigma_1^2 I_N\right),$$

où σ_1^2 est la composante "individuelle" de la variance, supposée inconnue.

(ii) ξ_2 est l'effet aléatoire temporel partagé par l'ensemble des individus, ce dernier pouvant être vu comme un phénomène latent non pris en considération dans les variables explicatives. Ayant tendance à perdurer au cours du temps, on le modélise à l'aide d'un processus autorégressif d'ordre 1 (AR(1)), i.e. pour tout $t \in \{1, \ldots, T-1\}$,

$$\begin{aligned} \xi_{2,t+1} &= \rho\, \xi_{2,t} + \nu_t, \\ \nu_t &\overset{\text{iid}}{\sim} \mathcal{N}(0,\, \sigma_2^2), \end{aligned}$$

où ρ est le paramètre de l'AR(1) et σ_2^2 la composante "temporelle" de la variance, supposés inconnus. De tels effets temporels latents apparaissent naturellement par exemple dans un contexte économique (où les agents partagent la même politique et conjoncture économiques dont les effets ont une certaine inertie temporelle), ou bien en biologie (car l'environnement écologique est souvent trop complexe pour être observé de manière exhaustive au travers des variables explicatives).

Enfin, ξ_1 and ξ_2 sont supposés indépendants. En notant $\xi = \left(\xi_1^T, \xi_2^T\right)^T$, $U_1 = I_N \otimes \mathbf{1}_T$, $U_2 = \mathbf{1}_N \otimes I_T$ and $U = [U_1 \,|U_2]$, le prédicteur linéaire η peut être écrit matriciellement :

$$\eta = X\beta + U\xi.$$

3 Méthodes

En raison de la structure de dépendance des GLMM, l'algorithme des scores de Fisher a été adapté par Schall (1991) afin d'estimer le modèle. Dans le but de tenir compte à la fois des fortes redondances dans X ainsi que des distributions non-conventionnelles des effets aléatoires, nous envisageons dans la suite la possibilité d'introduire une étape de type EM régularisé au sein d'un algorithme de Schall. Chaque itération se décompose alors en deux étapes clés : une étape de linéarisation et une étape d'estimation régularisée.

Étape de linéarisation. Pour tout $i \in \{1, \ldots, n\}$, la linéarisation à l'ordre 1 de y_i au voisinage de μ_i est donnée par : $g(y_i) \simeq z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$. Matriciellement, cette approximation fournit une variable dite "de travail" z s'exprimant au travers du modèle linéarisé suivant

$$\mathcal{M} : \quad z = X\beta + U\xi + e,$$

avec $\mathrm{Var}(e \,|\, \xi) = \mathrm{Diag}\left([g'(\mu_i)]^2 \,\mathrm{Var}(Y_i \,|\, \xi)\right)_{i=1,\ldots,n} = \Gamma$.

Étape d'estimation. Au lieu de résoudre le système de Henderson associé à $\mathcal{M}$ vu comme un LMM (à la manière de Schall (1991)), nous proposons plutôt une étape de type EM régularisé. Pour des données de petite dimension ($p < n$), nous suggérons d'étendre l'algorithme EM avec pénalité ridge élaboré par Eliot et al. (2011). Par contre, dans le cas $p \gg n$, nous lui préférons un algorithme EM régularisé basé sur la construction de composantes supervisées.

3.1 Données de petite dimension

Notre étape d'estimation prend appui sur Green (1990), qui popularise l'utilisation de l'algorithme EM lorsque la vraisemblance est pénalisée, et Golub et al. (1979), qui encouragent l'utilisation de la GCV pour choisir efficacement le coefficient de pénalisation λ. Cependant, contrairement au LMM homoscédastique considéré par Eliot et al. (2011), $\mathcal{M}$ contient des erreurs hétéroscédastiques. Nous optons alors plutôt pour le critère GCV proposé par Andrews (1991), p. 372, cohérent avec les modèles hétéroscédastiques. En notant $\theta = (\beta, \sigma_1^2, \sigma_2^2, \rho)$, nous présentons l'itération générique de notre algorithme EM pénalisé adapté aux GLMM avec effet aléatoire AR(1) dans l' Algorithme 1 ci-dessous.

3.2 Données de grande dimension

Dans le cas $p \gg n$, au lieu de maximiser une fonction objectif $\mathcal{Q}_{\text{pen}}$ pénalisée par la norme L_2 des coefficients, nous explorons la possibilité de maximiser une fonction $\mathcal{Q}_{\text{reg}}$ régularisée à l'aide de composantes. Pour une unique composante f, elle s'écrit sous la forme :

$$\begin{aligned} \mathcal{Q}_{\text{reg}}\left(\theta, \theta^{[t]}\right) &= \mathbb{E}_{\xi|z}\left[L_{\text{reg}}(\theta; z, \xi) \,|\, \theta^{[t]}\right], \text{ avec} \\ L_{\text{reg}}(\theta; z, \xi) &= (1-s)L(\theta; z, \xi) + s\phi(w), \end{aligned}$$

où $\phi(w)$ est un critère de pertinence structurelle (PS) introduit par Bry et Verron (2015) et $s \in [0, 1]$ un paramètre permettant de régler l'importance relative de la PS par rapport à L, vue ici comme une mesure de la qualité d'ajustemeent. Avec $l \geqslant 1$, $\phi(w)$ s'écrit :

$$\phi(w) = \left(\sum_{j=1}^{p} \left[\mathrm{cor}^2\left(x^j, f\right)\right]^l\right)^{\frac{1}{l}}.$$

Pour des raisons d'identifiabilité, la composante s'écrit $f = Cw$, avec $C = XU$ l'ensemble des composantes principales de X de valeurs propres non-nulles. Les paramètres s et l sont calibrés par validation croisée et les composantes de rangs supérieurs sont calculées comme celle de rang 1, après l'ajout de contraintes d'orthogonalité aux précédentes.

Algorithme 1 : Itération générique de l'algorithme EM pénalisé en norme L_2 pour GLMM avec effet aléatoire AR(1).

(1) **Étape de linéarisation.** Définir le modèle linéarisé par :

$$\mathcal{M}^{[t]} : z^{[t]} = X\beta + U\xi + e, \text{ avec } \mathrm{Var}(e \,|\, \xi) = \Gamma^{[t]}.$$

(2) **Étape d'estimation.**

(2.a) L désignant la log-vraisemblance complétée du modèle linéarisé, définir la log-vraisemblance complétée pénalisée L_{pen} par :

$$L_{\text{pen}}(\theta; z, \xi) := L(\theta; z, \xi) - \frac{\lambda}{2}\beta^T\beta$$

(2.b) Avec $\widehat{z}^{[t]}$ les valeurs ajustées et $S_\lambda^{[t]}$ la "hat-matrix" vérifiant l'égalité $\widehat{z}^{[t]} = S_\lambda^{[t]} z^{[t]}$, poser :

$$\lambda^{[t]} \longleftarrow \arg\min_\lambda \left\{ \mathrm{GCV}(\lambda) = \frac{n^{-1} \left\| z^{[t]} - S_\lambda^{[t]} z^{[t]} \right\|^2_{\Gamma^{[t]-1}}}{\left[1 - n^{-1}\mathrm{tr}\left(S_\lambda^{[t]}\right)\right]^2} \right\}.$$

(2.c) **Étape E.** Calculer :

$$\mathcal{Q}_{\text{pen}}\left(\theta, \theta^{[t]}\right) := \mathbb{E}_{\xi|z}\left[L_{\text{pen}}(\theta; z^{[t]}, \xi) \,|\, \theta^{[t]}, \lambda^{[t]}\right].$$

(2.d) **Étape M.** Poser alors :

$$\theta^{[t+1]} \longleftarrow \arg\max_\theta \mathcal{Q}_{\text{pen}}\left(\theta, \theta^{[t]}\right).$$

(3) **Mise à jour.** Poser $\xi^{[t+1]} = \mathbb{E}_{\xi|z}\left(\xi \,|\, \theta^{[t+1]}\right)$, et mettre à jour la variable de travail $z^{[t+1]}$ ainsi que la matrice de variance-covariance $\Gamma^{[t+1]}$.

Les étapes **(1)**–**(3)** sont répétées tant que la stabilité conjointe des paramètres β, σ_1^2, σ_2^2 et ρ n'est pas observée.

4 Résultats numériques

Afin d'évaluer les performances des deux méthodes, nous présenterons des études sur données simulées, notamment dans le cas Poisson - lien log. Ces simulations auront trois objectifs principaux :

(i) juger du nombre d'itérations nécessaire à la stabilisation des paramètres estimés, et ainsi se faire une idée de la vitesse de convergence des algorithmes proposés,

(ii) s'assurer que les MSE relatifs à chacun des paramètres convergent bien vers 0 lorsque la taille du jeu de données augmente,

(iii) vérifier que les méthodes proposées se comportent de manière identique quelle que soit la valeur de $\rho \in]-1, 1[$.

Bibliographie

[1] Andrews, D.W. (1991). Asymptotic optimality of generalized CL, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, **47**, 359 – 377.

[2] Bry, X., Trottier, C., Verron, T. et Mortier, F. (2013). Supervised component generalized linear regression using a pls-extension of the fisher scoring algorithm. *Journal of Multivariate Analysis*, **119**, 47 – 60.

[3] Bry, X. et Verron, T. (2015). THEME: THEmatic model exploration through multiple co-structure maximization. *Journal of Chemometrics*, **29**, 637 – 647.

[4] Chauvet, J., Trottier, C., Bry, X. et Mortier, F. (2016). Extension to mixed models of the Supervised Component-based Generalised Linear Regression. *COMPSTAT: Proceedings in Computational Statistics*.

[5] Eliot, M., Ferguson, J., Reilly, M.P. et Foulkes, A.S. (2011). Ridge Regression for Longitudinal Biomarker Data. *The International Journal of Biostatistics*, **7**, 1, Article 37.

[6] Golub, G.H., Heath, M. et Wahba, G. (1979). Generalized cross - validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215 – 223.

[7] Green, P.J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society, Series B*, **52**, 443 – 452.

[8] Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 – 727.

[9] Schelldorfer, J., Meier, L. et Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using l_1-penalization. *Journal of Computational and Graphical Statistics*, **23**, 460 – 477.