Bayesian high-dimensional variable selection in non-linear mixed-effects models using the SAEM algorithm

Marion Naveau, Guillaume Kon Kam King, Laure Sansonnet, Maud Delattre

Sélection de variables bayésienne en grande dimension dans les modèles non-linéaires à effets mixtes utilisant l'algorithme SAEM

Marion Naveau [1,2] & Guillaume Kon Kam King [2] & Laure Sansonnet [3] & Maud Delattre [2]

[1] *Université Paris-Saclay, INRAE, UMR MIA-Paris, 75005, Paris, France - E-mail : marion.naveau@inrae.fr*

[2] *Université Paris-Saclay, INRAE, UMR MaIAGE, 78350, Jouy-en-Josas, France E-mail : guillaume.kon-kam-king@inrae.fr, maud.delattre@inrae.fr*

[3] *Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris, 75005, Paris, France - E-mail : laure.sansonnet@agroparistech.fr*

Résumé. Les données de grande dimension, avec beaucoup plus de covariables que d'observations, comme les données génomiques par exemple, sont maintenant couramment analysées. Dans ce contexte, il est souvent souhaitable de pouvoir se concentrer sur les quelques covariables les plus pertinentes grâce à une procédure de sélection de variables. La question de la sélection de variables en grande dimension est largement documentée dans les modèles de régression standard, mais il existe encore peu d'outils pour y répondre dans le cadre des modèles à effets mixtes non linéaires. Dans ce travail, nous abordons la sélection de variables sous un angle bayésien et proposons une procédure de sélection combinant l'utilisation de priors *spike-and-slab* et l'algorithme SAEM. Comme pour la régression Lasso, l'ensemble des covariables pertinentes est sélectionné en explorant une grille de valeurs pour le paramètre de pénalisation. L'approche proposée est plus rapide qu'un algorithme MCMC classique et montre de très bonnes performances de sélection sur des données simulées.

Mots-clés. Sélection de variables bayésienne, modèles non linéaires à effets mixtes, données de grande dimension, *spike-and-slab*, algorithme SAEM.

Abstract. High-dimensional data, with many more covariates than observations, such as genomic data for example, are now commonly analyzed. In this context, it is often desirable to be able to focus on the few most relevant covariates through a variable selection procedure. High dimensional variable selection is widely documented in standard regression models, but there are still few tools to address it in the context of nonlinear mixed effects models. In this work, we approach variable selection from a Bayesian perspective and propose a selection procedure combining the use of *spike-and-slab* priors and the SAEM algorithm. Similarly to Lasso regression, the set of relevant covariates is selected by exploring a grid of values for the penalization parameter. The proposed approach is much faster than a classical MCMC algorithm and shows very good selection performances on simulated data.

Keywords. Bayesian variable selection, nonlinear mixed effects models, high-dimensional data, *spike-and-slab*, SAEM algorithm.

1 Introduction

1.1 Modèle statistique

On se place dans le contexte général d'observations collectées de façon répétée sur plusieurs individus au sein d'une population d'intérêt. On note n le nombre d'individus et n_i le nombre d'observations pour l'individu i. Sans perte de généralité et simplement pour alléger les notations, nous supposons que $n_i = J$ pour tout $i \in \{1, \dots n\}$. Les modèles non linéaires à effets mixtes (NLMEM) ont principalement été introduits pour modéliser les réponses d'individus ayant le même comportement global mais avec des variations individuelles ([Lavielle, 2014]). Pour tout $i \in \{1, \dots n\}$ et $j \in \{1, \dots J\}$, la réponse de l'individu i au temps t_{ij}, que l'on note y_{ij}, est modélisée de la façon suivante :

$$\begin{cases} y_{ij} = g(\varphi_i, t_{ij}) + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1a) \\ \varphi_i = \mu + {}^t\beta V_i + \xi_i & , \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Gamma^2) \quad (1b) \end{cases}$$

où g est une fonction connue qui dépend de manière non linéaire d'un paramètre individuel φ_i non-observé. Dans ce modèle, on suppose donc que toutes les réponses suivent une même fonctionnelle g mais dépendant de paramètres variables entre individus. Cette variabilité inter-individuelle est décrite au moyen de l'équation (1b) qui sépare, d'une part, la variabilité qui peut être expliquée par des covariables $V_i = {}^t(V_{i1}, \dots, V_{ip}) \in \mathbb{R}^p$, et d'autre part, la variabilité aléatoire, non attribuable aux covariables mesurées dans la population, elle-même décrite par l'effet aléatoire ξ_i.

Dans la suite, nous considérons les situations où $\varphi_i \in \mathbb{R}$. Nous notons $\theta = (\mu, \beta, \Gamma^2, \sigma^2)$ le paramètre de population, inconnu, $y = ((y_{ij})_{1 \leq j \leq J})_{1 \leq i \leq n}$ l'ensemble des observations disponibles et $\varphi = (\varphi_i)_{1 \leq i \leq n}$ l'ensemble des paramètres individuels.

1.2 Objectif et formulation bayésienne

L'objectif principal de ce travail est de sélectionner les covariables pertinentes, *i.e.* celles qui décrivent le mieux la variabilité entre les individus. Il s'agit donc d'estimer le support du vecteur d'effets fixes $\beta \in \mathbb{R}^p$, noté S^*_β :

$$S^*_\beta = \left\{ \ell \in \{1, \dots, p\} \middle| \beta_\ell \neq 0 \right\}$$

La difficulté ici est que le modèle (1) est un modèle à données incomplètes. En effet, bien que la première couche (1a) soit observée, ce n'est pas le cas pour les paramètres individuels $(\varphi_i)_i$. L'inférence est donc difficile car la vraisemblance et les estimateurs classiques n'ont pas de forme explicite.

L'estimation du paramètre de population θ dans les NLMEM a fait l'objet de nombreux travaux. Une solution très répandue est d'utiliser l'algorithme EM (*Expectation-Maximization*, [Dempster *et al.*, 1977]), ou une de ses variantes, comme l'algorithme SAEM

(*Stochastic Approximation EM*, [Delyon *et al.*, 1999]), pour calculer l'estimateur du maximum de vraisemblance ou l'estimateur du maximum *a posteriori*.

La sélection de variables dans les modèles à effets mixtes est elle aussi très étudiée. Les outils proposés sont très différents en petite dimension, *i.e.* $p < n$, et en grande dimension, *i.e.* $p >> n$, mais aussi selon que la fonction g est linéaire ou non linéaire en φ_i (voir par exemple [Müller *et al.*, 2013], [Delattre *et al.*, 2014], [Schelldorfer *et al.*, 2011], [Schelldorfer *et al.*, 2014]). Plus précisément, supposer g linéaire en φ_i dans (1) permet de développer des critères dont le calcul et/ou l'étude théorique font appel à des quantités explicites, ce qui est rarement le cas lorsque g est non linéaire en φ_i. En conséquence, la sélection de variables en grande dimension est peu traitée dans le cadre des NLMEM ([Bertrand et Balding, 2013], [Ollier, 2021]).

Dans ce travail, nous développons une nouvelle procédure de sélection de covariables en grande dimension dans les NLMEM. En s'inspirant des travaux de [Ročková et George, 2014] pour les modèles linéaires gaussiens, nous avons adopté une approche bayésienne utilisant un prior *spike-and-slab* pour le vecteur de régression β ([George et McCulloch, 1993]).

Pour cela, nous introduisons un vecteur binaire de variables latentes, $\delta = (\delta_\ell)_{1\leq\ell\leq p}$, correspondant au support de β, tel que :

$$\forall 1 \leq \ell \leq p \ , \ \delta_\ell = \begin{cases} 1 & \text{si la covariable } \ell \text{ est à inclure dans le modèle,} \\ 0 & \text{sinon.} \end{cases}$$

On peut remarquer qu'identifier les coefficients non nuls de β revient à identifier les composantes de δ égales à 1. Les priors bayésiens utilisés dans ce travail sont les suivants :

$$\begin{cases} \pi(\beta|\delta) & = \mathcal{N}_p(0, D_\delta), \text{ où } D_\delta = \text{diag}(a_1, \ldots, a_p),\ a_\ell = (1-\delta_\ell)\nu_0 + \delta_\ell\nu_1 \ , \ 0 < \nu_0 < \nu_1. & \text{(2a)} \\ \pi(\mu) & = \mathcal{N}(0, \sigma_\mu^2), \text{ avec } \sigma_\mu^2 > 0. & \text{(2b)} \\ \pi(\sigma^2) & = \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma\lambda_\sigma}{2}\right), \text{ avec } \nu_\sigma, \lambda_\sigma > 0. & \text{(2c)} \\ \pi(\Gamma^2) & = \mathcal{IG}\left(\frac{\nu_\Gamma}{2}, \frac{\nu_\Gamma\lambda_\Gamma}{2}\right), \text{ avec } \nu_\Gamma, \lambda_\Gamma > 0. & \text{(2d)} \\ \pi(\delta|\alpha) & = \alpha^{|\delta|}(1-\alpha)^{p-|\delta|}, \text{ avec } \alpha \in [0,1] \text{ et } |\delta| = \sum_{\ell=1}^p \delta_\ell. & \text{(2e)} \\ \pi(\alpha) & = Beta(a, b), \text{ avec } a, b > 0. & \text{(2f)} \end{cases}$$

où $(\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \nu_\Gamma, \lambda_\Gamma, a, b)$ sont des hyperparamètres fixés. Le prior fondamental est le prior *spike-and-slab* (2a) qui va induire la parcimonie dans le vecteur β. Une recommandation générale pour ce type de prior est de fixer l'hyperparamètre ν_0 petit et l'hyperparamètre ν_1 grand pour favoriser respectivement l'exclusion des effets non significatifs et l'inclusion des covariables influentes. Le choix des valeurs de ces hyperparamètres reste néanmoins difficile en pratique.

2 Procédure de sélection de variables

Nous proposons une procédure en deux étapes.

1. La première étape consiste à réduire le nombre de modèles candidats. De manière similaire à la régression Lasso, notre procédure explore une grille de valeurs pour la variance de la distribution spike ν_0 plutôt que de se concentrer sur une seule valeur. En effet, cet hyperparamètre contrôle le support de β, qui n'est pas connu. Ainsi, cette grille permet d'étudier différents niveaux de parcimonie dans le vecteur β. Notons Δ cette grille. Pour chaque $\nu_0 \in \Delta$, l'algorithme MCMC-SAEM, introduit par [Kuhn et Lavielle, 2004], est utilisé pour obtenir l'estimation par maximum *a posteriori* de $\Theta = (\theta, \alpha)$, notée $\widehat{\Theta}_{\nu_0}^{MAP}$. Celle-ci est ensuite utilisée pour obtenir une estimation $\widehat{\delta}$ des variables indicatrices δ en calculant la probabilité d'inclusion *a posteriori* de chaque covariable ℓ :

$$\widehat{\delta}_\ell = 1 \Longleftrightarrow \mathbb{P}(\delta_\ell = 1 | y, \widehat{\Theta}_{\nu_0}^{MAP}) \geq 0.5$$

On en déduit alors un sous-ensemble de covariables pertinentes :

$$\widehat{S}_{\nu_0} = \left\{ \ell \in \{1, \ldots, p\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP}) \right\}$$

où le seuil $s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP})$ est explicite et indépendant de la covariable ℓ considérée. Cette première étape réduit la collection de modèles à une collection de sous-modèles prometteurs $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$.

2. Ensuite, la deuxième étape consiste à sélectionner le "meilleur" modèle parmi ceux conservés à la fin de la première étape. On utilise une extension du BIC, le critère eBIC ([Chen et Chen, 2008]), dont la pénalité permet de tenir compte du fait que, comme $p >> n$, le nombre de modèles d'une certaine taille augmente très vite avec la taille considérée. Finalement le modèle sélectionné est le modèle $\widehat{S}_{\hat{\nu}_0}$ où :

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\operatorname{argmin}} \left\{ eBIC(\widehat{S}_{\nu_0}) \right\}$$

3 Applications numériques

Les performances de sélection de cette procédure ont été étudiées sur des données simulées selon un modèle de croissance logistique. Les résultats sont très encourageants, le bon support est sélectionné par notre procédure dans une grande majorité des cas. Comme attendu, pour différents nombres de covariables p fixés, le bon support est d'autant plus souvent sélectionné que le nombre d'individus n augmente et que la variance inter-individuelle Γ^2 diminue. Nous avons également l'intention de comparer notre méthode

à une procédure MCMC (Markov Chain Monte-Carlo) classique, et nous sommes assez confiants concernant la vitesse de notre méthode par rapport au MCMC.

Enfin, une application sur des données de sénescence de feuilles de blé est en cours. On s'intéresse à un stress azoté imposé à un panel de 220 variétés de blé tendre. Les variétés y répondent différemment, notamment on observe que certaines supportent mieux le stress et la sénescence est retardée. L'objectif est de sélectionner des marqueurs moléculaires, parmi plusieurs milliers, qui seraient associés à cette tolérance.

4 Remerciements

Ce travail a été financé par le projet Stat4Plant ANR-20-CE45-0012. Les auteurs remercient la plateforme INRAE MIGALE (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi : 10.15454/1.5572390655343293E12) pour les moyens de calcul et les capacités de stockage.

Bibliographie

[Bertrand et Balding, 2013] Bertrand, J. et Balding, D. J. (2013). Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenetics and genomics*, 23(3):167–174.

[Chen et Chen, 2008] Chen, J. et Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

[Delattre *et al.*, 2014] Delattre, M., Lavielle, M., Poursat, M.-A. *et al.* (2014). A note on bic in mixed-effects models. *Electronic journal of statistics*, 8(1):456–475.

[Delyon *et al.*, 1999] Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

[Dempster *et al.*, 1977] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1):1–22.

[George et McCulloch, 1993] George, E. I. et McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423): 881–889.

[Kuhn et Lavielle, 2004] Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8:115–131.

[Lavielle, 2014] Lavielle, M. (2014). *Mixed effects models for the population approach : models, tasks, methods and tools.* CRC press.

[Müller *et al.*, 2013] Müller, S., Scealy, J. L. et Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167. Publisher : Institute of Mathematical Statistics.

[Ollier, 2021] Ollier, E. (2021). Fast selection of nonlinear mixed effect models using penalized likelihood. *arXiv preprint arXiv :2103.01621*.

[Ročková et George, 2014] Ročková, V. et George, E. I. (2014). Emvs : The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

[Schelldorfer *et al.*, 2011] Schelldorfer, J., Bühlmann, P. et DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

[Schelldorfer *et al.*, 2014] Schelldorfer, J., Meier, L. et Bühlmann, P. (2014). Glmmlasso : an algorithm for high-dimensional generalized linear mixed models using 1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.