Efficient mode jumping MCMC for Bayesian variable selection in GLMM

Nord Stat 2016, June 27-30, 2016 Copenhagen, Denmark

Aliaksandr Hubin & Geir Storvik
Department of Mathematics, University of Oslo
aliaksah@math.uio.no, geirs@math.uio.no

UiO : Universitetet i Oslo

CELS

Abstract

Generalized linear mixed models (GLMM) are addressed for inference and prediction in a wide range of different applications providing a powerful scientific tool for the researchers and analysts coming from different fields. At the same time more sources of data are becoming available introducing a variety of hypothetical explanatory variables for these models to be considered. Estimation of posterior model probabilities and selection of an optimal model is thus becoming crucial. We suggest a novel mode jumping MCMC procedure for Bayesian model averaging and model selection in GLMM.

Introduction

In this study we address variable selection in generalized linear mixed models (GLMM) addressed in the Bayesian setting. These models allow to carry out detailed modeling in terms of both linking reasonably chosen responses and explanatory variables via a proper link function and incorporating the unexplained variability and dependence structure between the observations via random effects. Being one of the most powerful modeling tools in modern statistical science GLMM models have proven to be efficient in numerous applications from banking to astrophysics and genetics [2]. The posterior distribution of the models can be viewed as a relevant measure for the model evidence, based on the observed data. The number of models to select from is exponential in the number of candidate variables, moreover the search space in this context is often extremely non-concave. Hence efficient search algorithms have to be adopted for evaluating the posterior distribution of models within a reasonable amount of time. In this paper we introduce efficient mode jumping MCMC algorithms for calculating and maximizing posterior probabilities of the GLMM models.

Model and Inference

Generalized linear mixed models consist of a response Y_t coming from the exponential family distribution, a vector of P variables X_{ti} for observations $t \in \{1,...,T\}$ and latent indicators $\gamma_i \in \{0,1\}, i \in \{1,...,P\}$ defining if variable X_{ti} is included into the model ($\gamma_i = 1$) or not ($\gamma_i = 0$). We are also addressing the unexplained variability of the responses and the correlation structure between them through random effects δ_t with a specified parametric and sparse covariance matrix structure. Conditioning on the random effect we model the dependence of the responses on the explanatory variables via a proper link function $g(\cdot)$:

$$Y_t|\mu_t \sim \mathfrak{f}(y|\mu_t) \tag{1}$$
$$g(\mu_t) = \beta_0 + \sum_{i=1}^{P} \gamma_i \beta_i X_{ti} + \delta_t \tag{2}$$
$$\boldsymbol{\delta} = (\delta_1,...,\delta_T) \sim N_T(\mathbf{0}, \boldsymbol{\Sigma}_b). \tag{3}$$

Here $\beta_i \in \mathbb{R}, i \in \{0,...,P\}$ are regression coefficients showing in which way variables influence the linear predictor and $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\psi}) \in \mathbb{R}^T \times \mathbb{R}^T$ is the covariance structure of the random effect. We then put relevant priors for the parameters of the model in order to make a fully Bayesian inference:

$$\gamma_i \sim Binom(1,q) \tag{4}$$
$$\beta_i|\gamma_i \sim \mathbb{I}(\gamma_i = 1)N(\mu_\beta, \sigma_\beta^2) \tag{5}$$
$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}) \tag{6}$$

where q is the prior probability of including a covariate into the model.

Let $\boldsymbol{\gamma} = (\gamma_1,...\gamma_P)$, which uniquely defines a specific model. Then there are 2^P different fixed models in the space of models Ω_γ. We would like to find a set of the best models of this sort with respect to a certain model selection criterion - namely marginal posterior model probabilities (PMP) - $p(\boldsymbol{\gamma}|\mathbf{y})$, where $\mathbf{y}$ is the observed data. For the class of models addressed marginal likelihoods (MLIK) - $p(\mathbf{y}|\boldsymbol{\gamma})$ are obtained by the INLA approach [3]. Then PMP can be found using Bayes formula and estimated by iterating through the reasonable set of models $\mathbb{V}$ in the space of models Ω_γ.

$$p(\boldsymbol{\gamma}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \Omega_\gamma} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')} \approx \frac{\mathbb{I}(\boldsymbol{\gamma} \in \mathbb{V})p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathbb{V}} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}. \tag{7}$$

In (7) only models with high MLIK give significant contributions and thus iterating through them when constructing $\mathbb{V}$ is vital. The problem seems to be pretty challenging, because of both the cardinality of the discrete space Ω_γ growing exponentially fast with respect to the number of variables and the fact that Ω_γ is multimodal in terms of MLIK. Furthermore, the modes are often sparsely located [2]. For any other important parameters Δ the posterior distribution within our notation becomes

$$p(\Delta|\mathbf{y}) = \sum_{\boldsymbol{\gamma} \in \Omega_\gamma} p(\Delta|\boldsymbol{\gamma},\mathbf{y})p(\boldsymbol{\gamma}|\mathbf{y}), \tag{8}$$

whilst a model averaged expectation of a parameter Δ correspondingly is

$$\mathrm{E}[\Delta|\mathbf{y}] = \sum_{\boldsymbol{\gamma} \in \Omega_\gamma} \mathrm{E}[\Delta|\boldsymbol{\gamma},\mathbf{y}]p(\boldsymbol{\gamma}|\mathbf{y}). \tag{9}$$

Properties of the obtained in (7) - (9) estimators are also discussed in [2].

Mode Jumping MCMC

For generating the locally optimized proposals we first make a big jump to a new region of interest with respect to kernel $\mathsf{q}_\mathsf{I}(\boldsymbol{\chi}_0^*|\boldsymbol{\gamma})$, followed by some local optimization of $\pi(\boldsymbol{\gamma})$ with the chosen transition kernels $\mathsf{Q}_\mathsf{o}(\boldsymbol{\chi}_i^*|\boldsymbol{\chi}_{i-1}^*), i \in \{1,...,k\}$, which can be either stochastic or deterministic, and finally make randomization $\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}^*|\boldsymbol{\chi}_k^*)$ with a kernel based on a small neighborhood. For the reverse move we correspondingly first make a big jump $\mathsf{q}_\mathsf{I}(\boldsymbol{\chi}_0|\boldsymbol{\gamma}^*)$, followed by the same type of local optimization $\mathsf{Q}_\mathsf{o}(\boldsymbol{\chi}_i|\boldsymbol{\chi}_{i-1}), i \in \{1,...,k\}$, and finally the probability of transition from the point at the end of optimization to the initial solution γ is calculated with respect to the randomizing kernel $\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)$. A convenient choice of the auxiliary $h(\boldsymbol{\chi}|\boldsymbol{\gamma},\boldsymbol{\gamma}^*,\boldsymbol{\chi}^*)$ function [4] allowing to store very little of the information from the local optimization routine is to consider it of a form $h(\boldsymbol{\chi}|\boldsymbol{\gamma},\boldsymbol{\gamma}^*,\boldsymbol{\chi}^*) = \mathsf{h}(\boldsymbol{\chi}|\boldsymbol{\gamma},\boldsymbol{\gamma}^*)$:

$$\mathsf{h}(\boldsymbol{\chi}|\boldsymbol{\gamma},\boldsymbol{\gamma}^*) = \mathsf{q}_\mathsf{I}(\boldsymbol{\chi}_0|\boldsymbol{\gamma}^*)\left[\prod_{i=1}^{k} \mathsf{Q}_\mathsf{o}(\boldsymbol{\chi}_i|\boldsymbol{\chi}_{i-1})\right]. \tag{10}$$

Acceptance probabilities then reduce to:

$$r_m(\boldsymbol{\gamma},\boldsymbol{\gamma}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\gamma}^*)\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)}{\pi(\boldsymbol{\gamma})\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}^*|\boldsymbol{\chi}_k^*)}\right\}. \tag{11}$$

We recommend that in around $2-5\%$ of proposals mode jumping is performed for good mixing between the modes and accurate exploration of the regions around them. We address *accept the first improving neighbor*, *accept the best neighbor*, *simulated annealing*, and *local MCMC* approaches for performing local optimization, whilst transitions in these routines are based on random change or deterministic swaps of a fixed or randomized number of components of $\boldsymbol{\gamma}$, or by uniform addition or deletion of a positive component in $\boldsymbol{\gamma}$. Alternative MCMC estimators for (7) as described in [1, 2] are also available.

Results

We apply and compare the described algorithm further addressed as MJMCMC on the famous U.S. Crime Data (15 covariates) and the Protein Activity Data (88 covariates) and compare its performance to some popular algorithms such as BAS and competing MCMC methods (MC^3, RS, and thinned RS) with no mode jumping [1, 2].

U.S. Crime Data

We apply the Bayesian linear regression with a *g-prior* [1] to the aforementioned data sets with $T = 47$ observations and $P = 15$ explanatory variables. We carry out 100 replications of each algorithm on 10% of cardinality of Ω_γ, which in the best case scenario contains 86% of the total posterior model mass.

Parameter	Truth	MJMCMC		BAS	MC^3	RS	RS-thin
BIAS$\times 10^5$	0.00	15.49	9.28	10.94	27.33	27.15	27.3
RMSE$\times 10^5$	0.00	16.83	10.00	11.65	34.39	34.03	28.99
Explored mass	1.00	0.58	0.71	0.67	0.10	0.10	0.13
Unique models	32768	1909	3237	3276	829	1071	1722
Total models	32768	3276	5936	3276	3276	3276	3276

Table 1: BIAS, RMSE of posterior model probabilities, explored masses, total and efficient numbers of iterations from the 100 replications of the involved algorithms.

As can be seen from Table 1, our approach by far outperforms simpler MCMC methods in terms of the total posterior mass captured [1, 2] as well as the RMSE and BIAS [1, 2] of the model posterior probabilities (7); moreover, unlike the latter, it does not get stuck in the local modes and estimates a greater number of the unique models within the same amount of proposals. On the same amount of estimated models MJMCMC outperforms BAS in terms of all parameters, however for the same amount of proposals BAS is slightly better.

Protein Activity Data

Bayesian linear regression with a *g-prior* $T = 96$ observations and $P = 88$ explanatory variables is applied

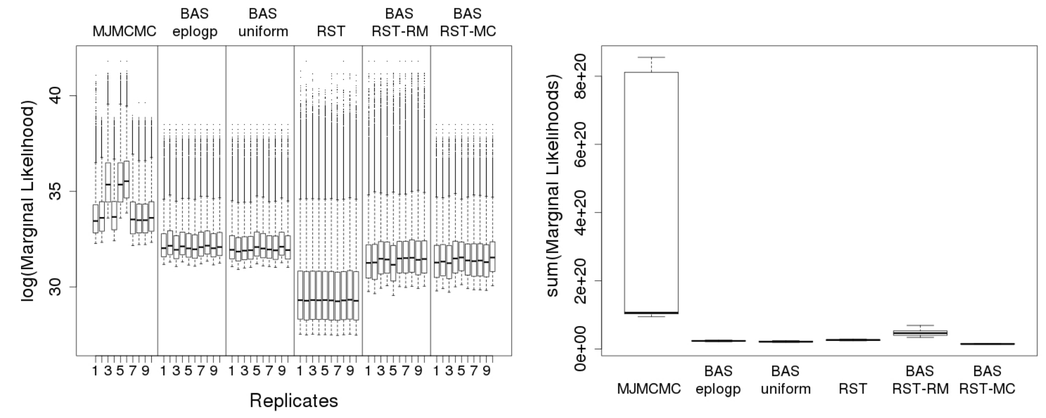

Figure 1: Comparions of the log marginal likelihood in the protein data of the top 100000 models (left) and boxplots of the posterior mass captured (right) obtained by MJMCMC, BAS-eplogp, BAS-uniform, thinned version of Random Swap (RST), BAS with Monte Carlo estimates of inclusion probabilities from the RST samples (BAS-RST-MC), and BAS renormalized estimates of inclusion probabilities (BAS-RST-RM) from the RST samples.

BAS with both uniform and eplogp initial sampling probabilities perform rather poorly in comparison to other methods, whilst BAS combined with RM approximations from RST as well as MJMCMC show the most promising results. BAS with RM initial sampling probabilities usually manages to find models with the highest MLIK, however MJMCMC in general captures by far higher posterior mass within the same amount of unique models addressed.

Conclusions

- Novel MJMCMC approach for estimating posterior model probabilities and Bayesian model averaging within GLMM and selection is introduced.
- MJMCMC incorporates the ideas of MCMC with possibility of large jumps combined with local optimizers to generate smart proposals in the discrete space of models
- *EMJMCMC* R-package is developed and available from the GitHub repository: `http://aliaksah.github.io/EMJMCMC2016` – simply scan the QR code on the top of the poster
- The developed package gives a user high flexibility in the choice of methods to obtain marginal likelihoods and model selection criteria within GLMM
- Extensive parallel computing for both MCMC moves and local optimizers is available within the developed package
- Based on the obtained results, MJMCMC can be claimed as a rather competitive novel algorithm in terms of the search quality.

Forthcoming Research

In future it would be of an interest to extend the procedure to the level of selection of link functions, priors and response distributions. The latter is expected to provide new horizons in automation of model selection and thus expand opportunities for addressing properly defined statistical models within machine learning applications. It will also require even more accurate tuning of parameters of the search introducing another important direction for further research.

References

[1] Merlise A. Clyde, Joyee Ghosh, and Michael L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011.

[2] Aliaksandr Hubin and Geir Storvik. Efficient mode jumping MCMC for Bayesian variable selection in GLMM, 2016. arXiv:1604.06398v2.

[3] Havard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Sosciety*, 71(2):319–392, 2009.

[4] Geir Storvik. On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation. *Scandinavian Journal of Statistics*, 38:342–358, 2011.

Acknowledgements

We would like to thank CELS project at the University of Oslo for giving us the opportunity, inspiration and motivation to write this article and Småforsk project for funding the conference participation.