Efficient mode jumping MCMC for Bayesian variable selection and model averaging in GLMM

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

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 GLMM addressed in the Bayesian setting. GLMM 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 sparse and non-concave. Hence efficient search algorithms have to be adopted for evaluating the posterior distribution of models within a reasonable amount of time. We introduce efficient mode jumping MCMC algorithm 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, 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 ($\gamma_i = 0$). The unexplained variability of the responses and the correlation structure between them are addressed via 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 put relevant priors for the parameters of the model in order to make a fully Bayesian inference:

$$\gamma_i \sim Binom(1,q) \tag{4}$$
$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_{P_\gamma}(\boldsymbol{\mu}_{\beta_\gamma}, \boldsymbol{\Sigma}_{\beta_\gamma}) \tag{5}$$
$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}) \tag{6}$$

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. 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 [4] or other methods [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.

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

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 growing exponentially fast with respect to the number of variables and the fact that Ω is multimodal in terms of MLIK. Furthermore, the modes are often sparsely located [2]. For any other important parameters Δ the model averaged distribution within our notation becomes:

$$p(\Delta|\mathbf{y}) = \sum_{\boldsymbol{\gamma} \in \Omega} p(\Delta|\boldsymbol{\gamma}, \mathbf{y})p(\boldsymbol{\gamma}|\mathbf{y}) \approx \sum_{\boldsymbol{\gamma} \in \mathbb{V}} p(\Delta|\boldsymbol{\gamma}, \mathbf{y})\tilde{p}(\boldsymbol{\gamma}|\mathbf{y}), \mathbb{V} \subseteq \Omega. \tag{8}$$

Properties of the obtained in (7) - (8) estimators are also discussed in [2].

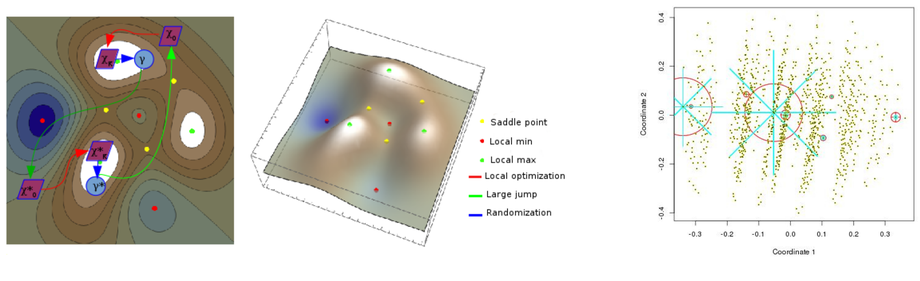

Figure 1: Illustration of locally optimized proposals (left) and MDS (multidimensional scaling) plot of the best 1024 models in terms of PMP in the space of models for some epigenetic data (right).

Mode Jumping MCMC

Locally optimized proposals are suggested at frequency ϱ of 2%-5% in order to increase the quality of proposals and consequently both improve the acceptance ratio and increase the probability of escaping from local optima (otherwise simple MCMC steps are performed). Assume $\boldsymbol{\chi}^* = (\boldsymbol{\chi}_0^*, \boldsymbol{\chi}_k^*)$ where $\boldsymbol{\chi}_0^*$ is generated according to some large jump proposal. $\boldsymbol{\chi}_0^*$ is modified to $\boldsymbol{\chi}_k^*$ through some optimization or simulation steps in order to move towards a local mode. Finally, $\boldsymbol{\gamma}^*$ is a randomized version of $\boldsymbol{\chi}_k^*$ such that irreducibility of the Markov chain is not violated. The procedure is illustrated in Figure 1 where the backward sequence $\boldsymbol{\gamma}^* \to \boldsymbol{\chi}_0 \to \boldsymbol{\chi}_k \to \boldsymbol{\gamma}$, needed for calculating the acceptance probability, is included.

1: **procedure** MODE JUMPING STEP
2: $\boldsymbol{\chi}_0^* \sim q_l(\cdot|\boldsymbol{\gamma})$ # make a large jump with kernel q_l
3: $\boldsymbol{\chi}_k^* \sim q_o(\cdot|\boldsymbol{\chi}_0^*)$ # perform local optimization with kernel q_o
4: $\boldsymbol{\gamma}^* \sim q_r(\cdot|\boldsymbol{\chi}_k^*)$ # randomize around the mode with kernel q_r
5: $\boldsymbol{\chi}_0 \sim q_l(\cdot|\boldsymbol{\gamma}^*)$ # make a reverse large jump with kernel q_l
6: $\boldsymbol{\chi}_k \sim q_o(\cdot|\boldsymbol{\chi}_0)$ # perform local optimization with kernel q_o
7: $r \leftarrow r_m(\boldsymbol{\chi}, \boldsymbol{\gamma}; \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\gamma}^*)q_r(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)}{\pi(\boldsymbol{\gamma})q_r(\boldsymbol{\gamma}^*|\boldsymbol{\chi}_k^*)}\right\}$ # calculate acceptance probability
8: **if** $\text{Unif}[0;1] \leq r$ **then**
9: $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}^*$ # accept the move to the proposed mode
10: **end if**
11: **end procedure**

Results

We apply the described algorithm further addressed as MJMCMC on 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]. We also report BMA predictions based on MJMCMC for the NEO asteroids data.

Protein Activity Data

Bayesian linear regression with a *g-prior* $T = 96$ observations and $P = 88$ explanatory variables is applied. Several approaches are compared on 20 replications with 1048576 models visited in each one.

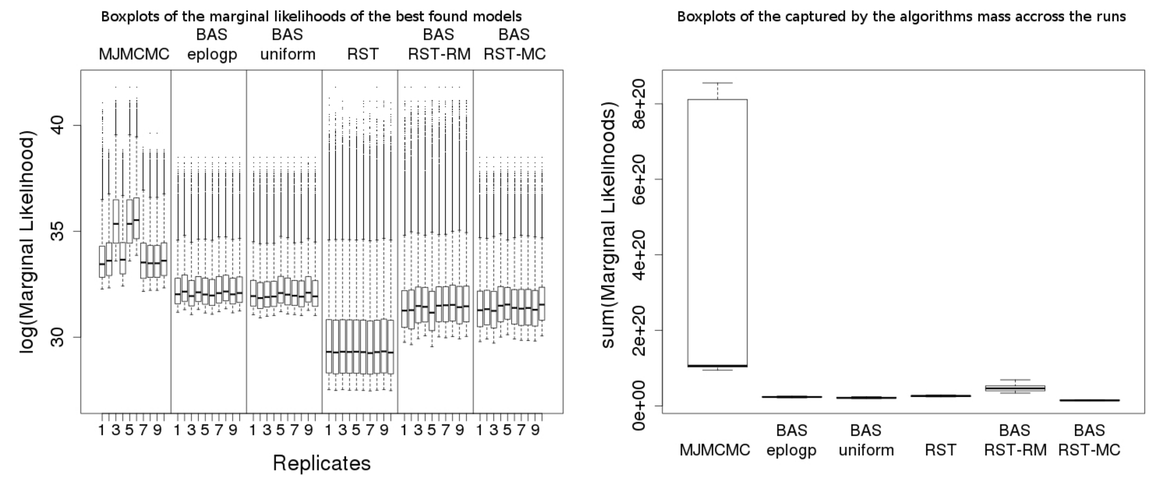

Figure 2: Comparisons of the log marginal likelihood in the protein data of the top 100000 models (left) and boxplots of the posterior mass captured (right) obtained by MJMCMC, BAS-eplogp, BAS-uniform, thinned version of Random Swap (RST), BAS with Monte Carlo estimates of inclusion probabilities from 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 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 visited.

NEO Asteroids Data

The observations are whether asteroids are potentially hazardous objects (PHA) or not (Phocaea). Logistic regression with an informative prior is applied. $P = 20$ different covariates describing objects are addressed. 64 objects are in the training set, 20720 - in the test set.

Subset	Cardinality	Accuracy	FNR	FPR
Ω	1048576	99.95656%	0.05670945%	0.01510117%
$\mathbb{V}$	20005	99.95656%	0.05670945%	0.01510117%
$\mathbb{V}$	10090	99.95656%	0.05670945%	0.01510117%
$\mathbb{V}$	2512	99.80212%	0.05670945%	0.49594239%
$\mathbb{V}$	412	99.46429%	0.04253813%	1.56110622%
$\mathbb{V}$	80	99.19402%	0.02836276%	2.40271201%
$\mathbb{V}$	4	90.00483%	0.04962427%	23.7651171%
$\mathbb{V}$	1	82.83301%	0.07087675%	34.8839473%

Table 1: Comparison of BMA predictive performance (Accuracy, FDR, FNR) based on MJMCMC and full screening on the test set

The results show high predictive opportunities of BMA predictions based on MJMCMC training. MJMCMC allows to avoid iterating through all 2^{20} models to obtain full predictive power. Though as we continue decreasing the number of models in $\mathbb{V}$, accuracy gradually drops.

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, inference and predictions.

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 as well as to allow feature engineering. This is expected to provide new opportunities in automation of model selection and thus expand capacity 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] M. Clyde, J. Ghosh, and M. Littman. Bayesian adaptive sampling for variable selection and model averaging. *J. Comp. Graph. Stat.*, 20(1):80–101, 2011.
[2] A. Hubin and G. Storvik. Efficient mode jumping MCMC for Bayesian variable selection in GLMM, 2016. arXiv:1604.06398v2.
[3] A. Hubin and G. Storvik. Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA), 2016. arXiv:1611.01450v1.
[4] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Royal Stat. Society*, 71(2):319–392, 2009.
[5] G. Storvik. On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation. *Scand. J. Stat.*, 38:342–358, 2011.

Acknowledgments

*The authors gratefully acknowledge the **CELS** project at the University of Oslo, `http://www.mn.uio.no/math/english/research/groups/cels/index.html`, for giving us the opportunity, inspiration and motivation to perform our research.*