ON MODE JUMPING IN MCMC FOR BAYESIAN VARIABLE SELECTION WITHIN GLMM

A. A. Hubin[1], G. O. Storvik[2]
University of Oslo
Oslo, NORWAY
e-mail: [1]aliaksah@math.uio.no, [2]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.

1 Introduction

In this paper we study 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, 3]. 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.

2 The generalized linear mixed regression model

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 $g(\cdot)$:

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

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:

$$\gamma_i \sim Binom(1, q),\ \beta_i|\gamma_i \sim \mathbf{1}(\gamma_i = 1)N(\mu_\beta, \sigma_\beta^2),\ \boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}),$$

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 [5]. 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.

$$p(\boldsymbol{\gamma}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \Omega_\gamma} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')} \approx \frac{\mathbf{1}(\gamma \in \mathbb{V})p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathbb{V}} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')}. \quad (1)$$

In (1) 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 $\Omega_{\boldsymbol{\gamma}}$ growing exponentially fast with respect to the number of variables and the fact that $\Omega_{\boldsymbol{\gamma}}$ is multimodal in terms of MLIK. Furthermore, the modes are often sparsely located [3]. [3] also report and discuss properties of the obtained in (1) estimator.

3 Mode jumping MCMC

In the MCMC approach as described by [4], Metropolis-Hastings algorithms are addressed as a class of methods for drawing from a complicated target distribution. [6] describes high potential flexibility in choices of proposals by means of generating additional auxiliary states allowing cases where the proposal densities are not directly available. The auxiliary states can be chains generated by some local optimizers chosen randomly from a mixture and allowing for jumps to alternative modes. [6] shows that the detailed balance equations is satisfied for this general case. Assume the current state to be $\boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma})$. Generate $(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*) \sim q(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*|\boldsymbol{\gamma})$ and consider $\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^* \sim h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*)$ as some auxiliary variables for some arbitrary chosen $h(\cdot|\cdot)$. Accept $\boldsymbol{\gamma}' = \boldsymbol{\gamma}^*$ with the following acceptance probability

$$r_m(\boldsymbol{\chi}, \boldsymbol{\gamma}; \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\gamma}^*)h(\boldsymbol{\chi}^*|\boldsymbol{\gamma}^*, \boldsymbol{\chi}, \boldsymbol{\gamma})q(\boldsymbol{\chi}, \boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{\pi(\boldsymbol{\gamma})h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\chi}^*, \boldsymbol{\gamma}^*)q(\boldsymbol{\chi}^*, \boldsymbol{\gamma}^*|\boldsymbol{\gamma})}\right\}, \quad (2)$$

or remain in the previous state otherwise. Then an ergodic Markov chain is generated and $\boldsymbol{\gamma}' \sim \pi(\boldsymbol{\gamma}')$. In a typical setting $\boldsymbol{\chi}^*$ is generated first, followed by $\boldsymbol{\gamma}^*$. The extra $\boldsymbol{\chi}$ is needed in order to calculate a legal acceptance probability, relating to a symmetric reverse move.

For generating the locally optimized proposals we first make a big jump to a new region of interest with respect to kernel $\mathsf{q}_\mathsf{l}(\boldsymbol{\chi}_0^*|\boldsymbol{\gamma})$, 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 $\mathsf{q}_\mathsf{l}(\boldsymbol{\chi}_0|\boldsymbol{\gamma}^*)$, 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 $\mathsf{q}_\mathsf{r}(\boldsymbol{\gamma}|\boldsymbol{\chi}_k)$. Then acceptance probabilities with respect to (2) are calculated and the move to a new state is either accepted or rejected. A convenient choice of $h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \boldsymbol{\chi}^*)$ function allowing to store very little of the information from the local optimization routine is to consider it of a form $h(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*, \boldsymbol{\chi}^*) = \mathsf{h}(\boldsymbol{\chi}|\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$:

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

Then (2) reduces to

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

We recommend that in not less than 95% of the proposals no mode jumping is performed. This provides the global Markov chain with both good mixing between the modes and accurate exploration of the regions around them. As described by [3] we address *accept the first improving neighbor*, *accept the best neighbor*, *simulated annealing*, and *local MCMC* approaches for performing local combinatorial 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 $\boldsymbol{\gamma}$. Notice that tuning of the probabilities of addressing local optimizers with particular proposal kernels in a mixture is often beneficial and we can carry it out during the burn in of the mode jumping MCMC without violating the desired ergodicity of the chain [3]. Also notice that both local optimizers and the global MCMC procedures are extensively parallelizible [3]. Finally, all of the unique models visited during the procedure are then appended to $\mathbf{V} \subseteq \Omega_\gamma$ and used to estimate (1). Alternative MCMC estimators for (1) as described in [1, 3, 4] are also available.

4 Results and discussion

We apply and compare the described algorithm further addressed as MJMCMC on the famous U.S. Crime Data 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, 3]. We apply the Bayesian linear regression with a *g-prior* [1] to the aforementioned data set 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. As can be seen

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.

from Table 1, our approach by far outperforms simpler MCMC methods in terms of the total posterior mass captured [1, 3] as well as the RMSE and BIAS [1, 3] of the model posterior probabilities (1); 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. More examples with various GLMM addressed and description of the developed R package *EMJMCMC* can be found in [3]. In general, we claim that MJMCMC is not only a very competitive novel algorithm, but also that it addresses a much wider class of models (GLMM) than all of the competing approaches. In future it would be of an interest to extend the procedure to level of the choice of link functions, priors and response distributions.

References

[1] Clyde M., Ghosh J., Littman M. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comp. Graph. Stat.* Vol. **20**(1), pp. 80–101.

[2] Cressie N., Wikle C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, NJ.

[3] Hubin A., Storvik G. (2016). Efficient mode jumping MCMC for Bayesian variable selection in GLMM. `arXiv:1604.06398v1`

[4] Robert C., Casella G. (2005). *Monte Carlo statistical methods*. Springer, NY.

[5] Rue H., Martino S., Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Royal Statistical Sosciety*. Vol. **71**(2), pp. 319–392.

[6] Storvik G. (2011). On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation. *Scand. J. Stat.* Vol. **38**, pp. 342–358.