Efficient mode jumping MCMC for Bayesian variable selection in GLMM

Hubin A.A.,Storvik G.O.

Department of Mathematics, University of Oslo

aliaksah@math.uio.no, *geirs@math.uio.no*

UiO **:** Universitetet i Oslo

UiO, Department of Astrophysics, Oslo

30.05.2016

Introduction. Issues

- GLMM are addressed for inference and prediction in a wide range of different applications providing a powerful scientific tool for the researchers and analysts from different fields

- More and more sources of data are becoming available introducing a variety of hypothetical explanatory variables for these models to be considered

- Selection of an optimal combination of these variables is crucial. Posterior model probabilities is one of the relevant measures to estimate quality of the models

- The number of models to select from is exponential in the number of candidate variables

- The search space in this context has numerous local extrema (potentially sparsely located)

- Hence efficient search algorithms have to be adopted for evaluating the posterior distribution within a reasonable amount of time

Bayesian Generalized Linear Mixed Model

$$Y_t|\mu_t \sim f(y|\mu_t), t \in \{1, ..., T\} \tag{1}$$

$$\mu_t = g^{-1}(\eta_t) \tag{2}$$

$$\eta_t = \gamma_0\beta_0 + \sum_{i=1}^{p} \gamma_i\beta_i X_{ti} + \delta_t \tag{3}$$

$$\boldsymbol{\delta} = (\delta_1, ..., \delta_T) \sim N_T(\mathbf{0}, \boldsymbol{\Sigma}_b). \tag{4}$$

- $\beta_i \in \mathbb{R}, i \in \{0, ..., p\}$ are regression coefficients
- $\boldsymbol{\Sigma}_b = \boldsymbol{\Sigma}_b(\boldsymbol{\psi}) \in \mathbb{R}^T \times \mathbb{R}^T$ is the covariance of the random effect δ_t
- $g(\cdot)$ is a proper link function
- $\gamma_i \in \{0, 1\}, i \in \{0, ..., p\}$ are latent indicators defining if covariate X_{ti} is included into the model ($\gamma_i = 1$) or not ($\gamma_i = 0$)

We use a fully Bayesian approach, hence specify priors

$$\gamma_i \sim Binom(1, q) \tag{5}$$

$$q \sim Beta(\alpha_q, \beta_q) \tag{6}$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_{\sum_{i=1}^{p} \gamma_i}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \tag{7}$$

$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}), \tag{8}$$

- q is the prior probability of including a covariate into the model
- α_q, β_q are hyper parameters for the prior on q
- $\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta$ are hyper parameters for the prior on $\boldsymbol{\beta}|\boldsymbol{\gamma}$
- $\boldsymbol{\psi}$ are the hyper parameters of the random effect

Inference on the model

Let:

$\theta = \{\vec{\beta}, \rho, \sigma_\epsilon^2\}$ define parameters of the model and $\gamma : \vec{\gamma}$ define a model itself, i.e. which covariates are addressed.

Then:

- $\theta|\gamma$ define parameters conditioned on fixed models
- $\exists 2^{p+1}$ different models

Goals:

- $p(\gamma, \theta|\mathbb{D})$ posterior distribution of parameters and models
- $p(\gamma|\mathbb{D})$ marginal posterior distribution of the models
- Set of estimated models performing well in terms of some model selection criteria (MAP, WAIC, DIC, MLIK)

Procedure

- **Note that** $p(\gamma, \theta | \mathbb{D}) = p(\theta | \gamma, \mathbb{D}) p(\gamma | \mathbb{D})$
- $p(\theta | \gamma, \mathbb{D})$ and $\log p(\mathbb{D} | \gamma)$ can be efficiently obtained by INLA
- **Note that** $p(\gamma | \mathbb{D}) = \dfrac{e^{\log p(\mathbb{D} | \gamma) + \log p(\gamma)}}{\sum_{\gamma' \in \Omega_\gamma} e^{\log p(\mathbb{D} | \gamma') + \log p(\gamma')}}$
- $\widehat{p}(\gamma | \mathbb{D}) = \dfrac{e^{\log p(\mathbb{D} | \gamma) + \log p(\gamma)}}{\sum_{\gamma' \in \mathbb{V}} e^{\log p(\mathbb{D} | \gamma') + \log p(\gamma')}}$
- $\mathbb{V}$ is the subspace of Ω_γ to be efficiently explored
- Note that for $p(\gamma) = p(\gamma') \forall \gamma, \gamma' \in \Omega_\gamma$:
- $p(\gamma | \mathbb{D}) \gg p(\gamma' | \mathbb{D})$ if $\log p(\mathbb{D} | \gamma) > \log p(\mathbb{D} | \gamma')$ often $\implies$
- **Near modal values in terms of log MLIK are particularly important** for construction of reasonable $\mathbb{V} \subset \Omega_\gamma$, **missing** them **can dramatically influence** posterior in the original space Ω_γ

Main challenges are multimodality in Ω_γ and its size.

- Full enumeration of Ω_γ - infeasible for large dimensions
- Random walk in Ω_γ including simple MCMC - does not take advantage of the structure of $\Omega_\gamma \Longrightarrow$ too slow
- Greedy optimization with numerous initial points - end up in local optima
- Random walk with mode jumping proposals seems to be a good idea

Search Space
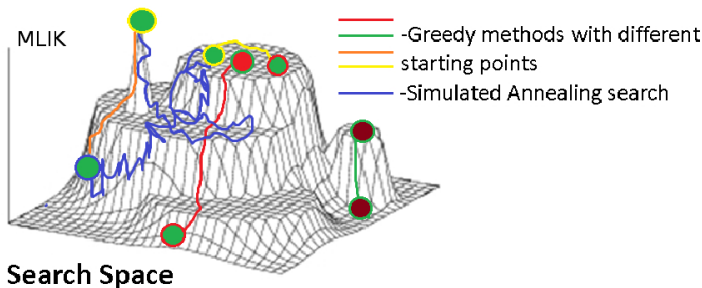
MCMC with locally optimized proposals

Tjelmeland and Hegstad [6] suggested continuous mode jumping proposals, **Storvik** [5] considers a more general setup, **we suggest mode jumping proposals** in the **discrete parameter** spaces.
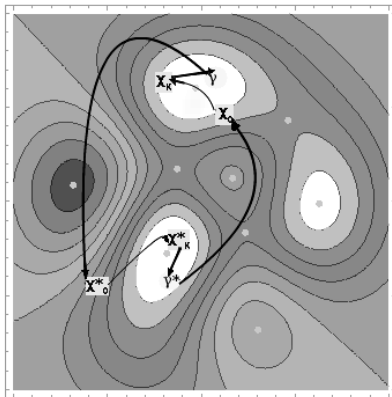

Figure: Locally optimized with randomization proposals

Application of MCMC with mode jumping proposals

We have shown that the detailed balance equation is satisfied for the following acceptance probabilities:

$$r_m(\gamma_j, \gamma_k) = \min \left\{ 1, \frac{p(\mathbb{D}|\gamma_k)p(\gamma_k)q_s(\gamma_j|\gamma_{j_{K-1}})}{p(\mathbb{D}|\gamma_j)p(\gamma_j)q_s(\gamma_k|\gamma_{k_{K-1}})} \right\}. \tag{9}$$

- $q_s(.|.)$ is the kernel of randomization at the end.

Hence we also obtain alternative MCMC estimators of posterior marginal probabilities

$$\tilde{p}(\gamma|\mathbb{D}) = \frac{\sum_{i=1}^{W} \mathbb{I}(\gamma_i = \gamma)}{W} \xrightarrow[W \to \infty]{d} p(\gamma|\mathbb{D}). \tag{10}$$

- W is the number of MCMC iterations (after burn-in)

How it looks like in reality

Modes are important: the standard MCMC procedure (right) misses two in this example. Visualization is challenging
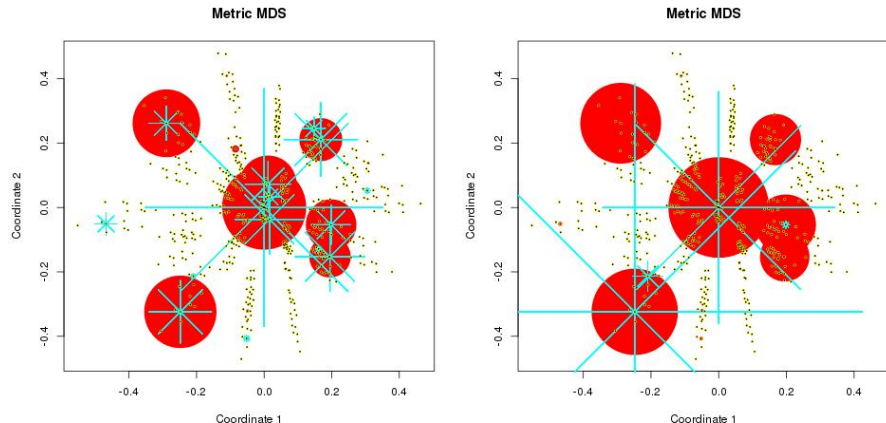

Figure: MDS plots with posterior modes of all found solutions for the approaches

Application to cosmological simulations. Cosmological hydro-simulation data (http://yt-project.org/data/).

Observations (Bernoulli classifiers):

- Galaxy is quenched (or not)
- Star hosts planet (or not)

Variables:

- Dark matter mass
- Gas mass
- Stellar mass
- Star formation rate
- Metallicity
- Gas molecular fraction
- Gas fraction
- Stellar fraction
- Stellar to gas mass ratio
- Other covariates, their interactions, polynomes and etc.

Application to NEO classification. NASA Space Challenge (`https://github.com/SpaceApps2016/Resources`).

Observations (Bernoulli classifiers):

- Asteroid is a NEO (PHA) object or not (Phocaea)

Variables:

- Rotation period
- Magnitude slope
- Mean anomaly
- Inclination
- Argument of perihelion
- Longitude of the ascending node
- Rms residual
- Semi major axis
- Eccentricity
- Mean motion
- Absolute magnitude
- Other covariates, their interactions, polynomes and etc.

Application to cosmological simulations or NEO objects classification
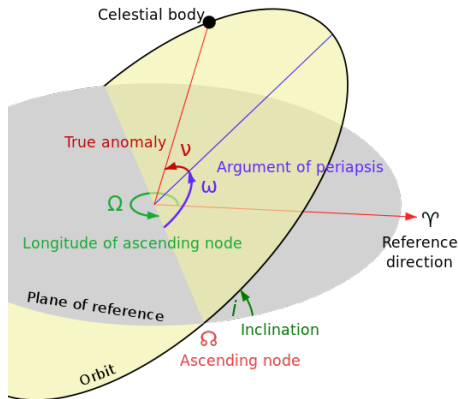
Logistic Bayesian regression addressed

$$y_t = y | p_t \sim Binom(1, p_t) \tag{11}$$

$$p_t = \frac{e^{\gamma_0 \beta_0 + \sum_{i=1}^{P} \gamma_i \beta_i x_{t,i}}}{1 + e^{\gamma_0 \beta_0 + \sum_{i=1}^{P} \gamma_i \beta_i x_{t,i}}} \tag{12}$$

$$\boldsymbol{\beta} | \boldsymbol{\gamma} \sim N_{\sum_{i=1}^{P} \gamma_i}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \tag{13}$$

$$\gamma_i \sim Binom(1, q) \tag{14}$$

Figure: Orbital elements(left) by Lasunncty (talk), CC BY-SA 3.0 and absolute vs apparent magnitude (right) by Mrscreath(http://mrscreath.blogspot.com)

20 covariates addressed in the experiment (both *reasonable* and *heuristic*): Mean anomaly $\in [0°; 360°)$**;** Argument of perihelion $\in [0°; 360°)$**;** Longitude of the ascending node $\in [0°; 360°)$**;** Inclination $\in [0°; 180°]$**;** Semi major axis $\in \mathbf{R}^+$**;** Eccentricity $\in \mathbf{R}^+$**;** Mean motion $\in \mathbf{R}^+$**;** Absolute magnitude $\in \mathbf{R}$ (brightness)**;** Rms residual $\in \mathbf{R}^+$ (brightness error)**;** Eccentricity$^2 \in \mathbf{R}^+$**;** Absolute magnitude$^2 \in \mathbf{R}^+$**;** Semi major axis$^2 \in \mathbf{R}^+$**;** Semi major axis$^3 \in \mathbf{R}^+$**;** Mean anomaly×Semi major axis**;** Mean anomaly×Semi major axis$^2 \in \mathbf{R}^+$**;** Mean anomaly×Semi major axis$^3 \in \mathbf{R}^+$**;** Argument of perihelion×Semi major axis $\in \mathbf{R}^+$**;** Argument of perihelion×Semi major axis$^2 \in \mathbf{R}^+$**;** Argument of perihelion×Semi major axis$^3 \in \mathbf{R}^+$**;** Longitude of the ascending node×Semi major axis $\in \mathbf{R}^+$.

Training set includes 32 NEO and 32 non-NEO objects, **test set** includes 20720 objects (14099 NEO, 6621 non-NEO), **validation sets** were used as some random subsets of a 100 elements from these 20720 **objects**

2^{20} models in total, algorithm was run until ca **2500 models** and ca **10000 models** are visited.

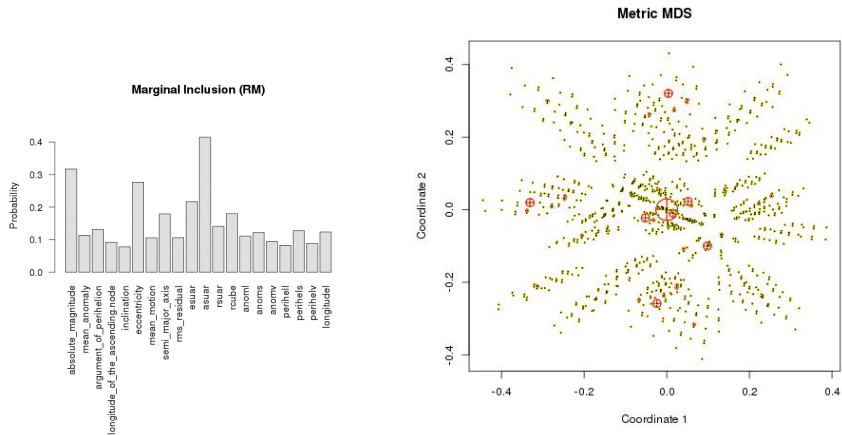Posterior inclusion probabilities and posterior model probabilities

Figure: Comparison of marginal inclusion probabilities of the covariates (left) and models on the whole (right)

NEO objects classification. Bayesian classification

Choice of $\mathbb{V}^*$ is crucial, $\mathbb{V}^* = \Omega_\gamma$ - often in-feasible, $\mathbb{V}^* = \mathbb{V}$ - very precise can be too slow, $\mathbb{V}^* = \mathbb{V} \cap p(\gamma|\mathbb{D}) \geq \alpha$ - often precise, but is a way faster!!!
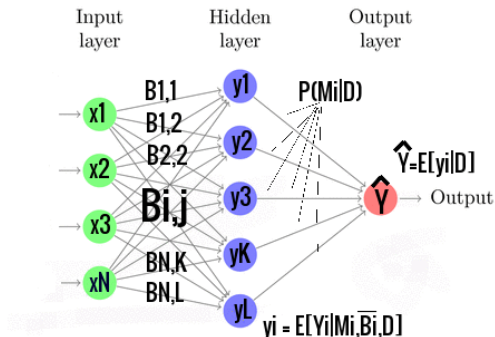

Figure: Bayesian Artificial Neuron Network for Classification

$$\hat{Y} = \mathbb{I}\{\hat{E}[Y|\mathbf{D}] \geq 0.5\}, \hat{E}[Y|\mathbf{D}] = \sum_{\gamma \in \mathbb{V}^*} \hat{E}[y_\gamma|\gamma, \mathbf{D}]\hat{p}(\gamma|\mathbf{D})$$

NEO objects classification. Results

Quite impressive actually... Surprisingly or not?.. Comments?..
Remember: $\|\textbf{training set}\| = 64$, $\|\textbf{test set}\| = 20720$

Subset	$\|$Hidden$\|$	Precision	FNR	FPR	Time	Time/it.
$\mathbb{V}^1$	10090	99.95656%	0.05670945 %	0.01510117 %	619.89 min	1.795 sec
$\mathbb{V}^2$	2512	99.80212%	0.05670945 %	0.49594239 %	172.66 min	0.499 sec
$\mathbb{V}^3$	412	99.46429%	0.04253813 %	1.56110622%	29.166 min	0.084 sec
$\mathbb{V}^4$	80	99.19402%	0.02836276%	2.40271201%	9.9812 min	0.029 sec
$\mathbb{V}^5$	4	90.00483%	0.04962427 %	23.7651171 %	4.7789 min	0.014 sec
$\text{argmax}_{\gamma \in \mathbb{V}^1} \{ p_{\mathbb{V}}(\boldsymbol{\gamma}|\mathbb{D}) \}$	1	82.83301%	0.07087675 %	34.8839473 %	4.5222 min	0.013 sec
Wake up NEO	?	93.86271%	1.00000000%	17.0000000%	-	-

Table: Comparison of performance (Precision, FDR, FNR, Time) of different models

N/B: the best model includes eccentricity^2, eccentricity, absolute magnitude^2, absolute magnitude
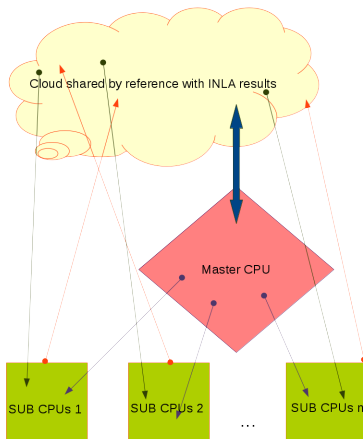
Figure: Multiprocessing architecture

Comparison to other algorithms: BAS, RS (simper MCMC) on 2^{20} unique models visited for MJMCMC and BAS and 88×2^{20} iterations of RS.
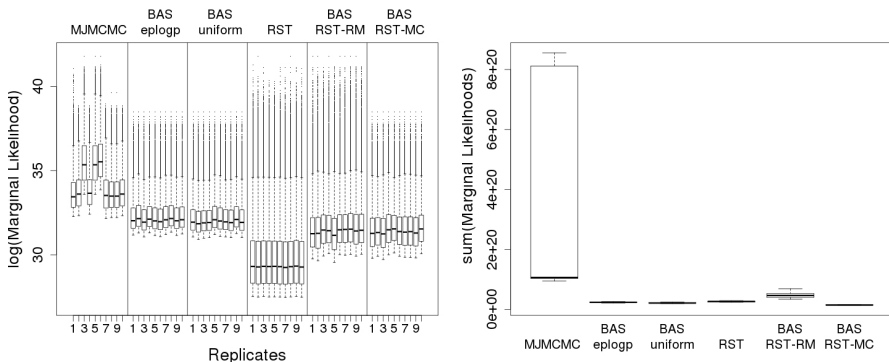

Figure: 100000 best mliks found (left) and posterior masses captured (right). Bayesian linear regression with a g-prior is addressed, since no other packages (to our awareness) manage model selection in GLMM
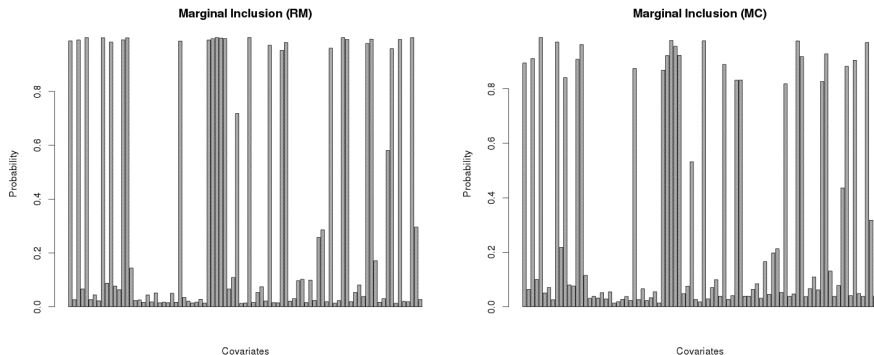
Checking convergence. Marginal inclusion probabilities

Figure: Comparison of marginal inclusion probabilities obtained by the Bayes formula and MCMC approximations from the best run of MJMCMC with 8.56e + 20 posterior mass captured

Concluding remarks

- We introduced the MJMCMC approach for estimating posterior model probabilities and Bayesian model averaging and selection.
- It incorporates the ideas of MCMC with possibility of large jumps combined with local optimizers to generate proposals in the discrete space of models
- *EMJMCMC* R-package is developed and available from the GitHub repository: `http://aliaksah.github.io/EMJMCMC2016/`
- 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 in the experimental part results, we can claim MJMCMC to be a rather competitive novel algorithm that both performs well in terms of the search quality and addressed a more general class of statistical models than the competing approaches

References

📄 M. Clyde, J. Ghosh, and M. Littman.
Bayesian adaptive sampling for variable selection and model averaging.
Journal of Computational and Graphical Statistics, 20(1):80–101, 2011.

📄 A. Hubin and G.O. Storvik
Efficient mode jumping MCMC for Bayesian variable selection in GLMM.
arXiv:1604.06398v1, 2016.

📄 H. Rue, S. Martino, and N. 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.

📄 G.O. Storvik.
On the flexibility of metropolis-hastings acceptance probabilities in auxiliary variable proposal generation.
Scandinavian Journal of Statistics, 38:342–358, 2011.

The End.

Thank you.