On mode jumping in MCMC for Bayesian variable selection within 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

11th International Conference
COMPUTER DATA ANALYSIS AND MODELING
Minsk 2016

07.06.2016

Introduction. Issues

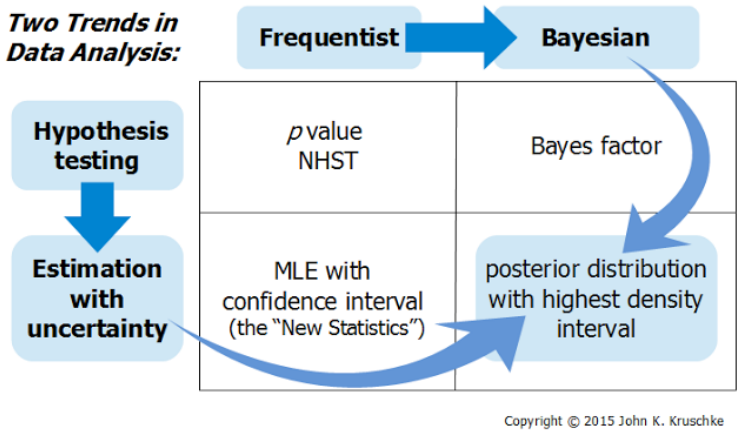
- GLMM are used in a wide range of different applications for
  - Inference
  - Predicton
- More sources of data → more hypothetical explanatory variables →
  - Model selection
  - Model averaging
- Posterior marginal model probabilities are used to
  - Estimate quality of the models
  - Serve as weights in Bayesian model averaging
- Eficient search algorithms have for evaluating posterior marginal model probabilities are required since
  - The number of models to select from is exponential in the number of candidate variables
  - The search space has numerous sparcely located local extrema
  - Time and computing resources are limited

Bayesian vs. Frequentist statistics

Frequentist: obtain $\hat{\theta}$ with CI by MLE, MM, MD etc.

vs.

Bayesian: obtain $p(\theta|\mathbb{D}) = \frac{p(\mathbb{D}|\theta)p(\theta)}{p(\mathbb{D})} = \frac{p(\mathbb{D}|\theta)p(\theta)}{\int_{\Omega_{\theta'}} p(\mathbb{D}|\theta')p(\theta')d\theta'}$



Frequentist vs. Bayesian

Frequentist	Bayesian
Probability is a long-run average	Probability is a degree of belief
There is a true Model, the Data is a random realization	The Data is true/fixed, Models have probabilities
Probability of the data given a hypothesis (Likelihood)	Probability of a hypothesis given the data
Each repeated experiment/observation starts from ignorance	Can incorporate prior knowledge: probabilities can be updated


Jerzy Neyman


Harold Jeffereys

Figure: Paradigms shifts (left, adopted from John K. Kruschke, `doingbayesiandataanalysis.blogspot.no`) and differences between the paradigms (right, adopted from Andres Lopez-Sepulcre, `www.slideshare.net/andreslopezsepulcre`)

Bayesian Generalized Linear Mixed Model

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

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

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

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

- $\beta_i \in \mathbb{R}, i \in \{0, ..., p\}$ 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

$$\gamma_i \sim Binom(1, q) \tag{5}$$
$$q \sim Beta(\alpha_q, \beta_q) \tag{6}$$
$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), u = \sum_{i=1}^{p} \gamma_i \tag{7}$$
$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}). \tag{8}$$

- 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 $\boldsymbol{\beta}|\boldsymbol{\gamma}$
- ψ are the hyper parameters of the random effect

Inference on the model

Let:

- $\boldsymbol{\gamma} = \vec{\gamma}$ define a model itself, i.e. which covariates are addressed
- θ define parameters of the model

Then:

- $\exists 2^{p+1}$ different models

Goals:

- $p(\boldsymbol{\gamma}, \theta | \mathbb{D})$ posterior distribution of parameters and models
- $p(\boldsymbol{\gamma} | \mathbb{D})$ marginal posterior probabilities of the models
- $p(\mathfrak{S} | \mathbb{D})$ marginal posterior probabilities of the quantiles of interest $\mathfrak{S}$

Procedure

- **Notice that** $p(\gamma, \theta|\mathbb{D}) = p(\theta|\gamma, \mathbb{D})p(\gamma|\mathbb{D})$
- $p(\theta|\gamma, \mathbb{D})$ and $p(\mathbb{D}|\gamma)$ can be efficiently obtained by INLA
- **Notice that** $p(\gamma|\mathbb{D}) = \frac{p(\mathbb{D}|\gamma)p(\gamma)}{\sum_{\gamma' \in \Omega_\gamma} p(\mathbb{D}|\gamma')p(\gamma')}$
- **Approximate with** $\widehat{p}(\gamma|\mathbb{D}) = \frac{p(\mathbb{D}|\gamma)p(\gamma)}{\sum_{\gamma' \in \mathbb{V}} p(\mathbb{D}|\gamma')p(\gamma')}$
- $\mathbb{V}$ is the **subspace** of Ω_γ to be **efficiently explored**
- **Near modal values in terms of 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

Problems with standard MCMC

Main challenges are multimodality in Ω_γ and its size.

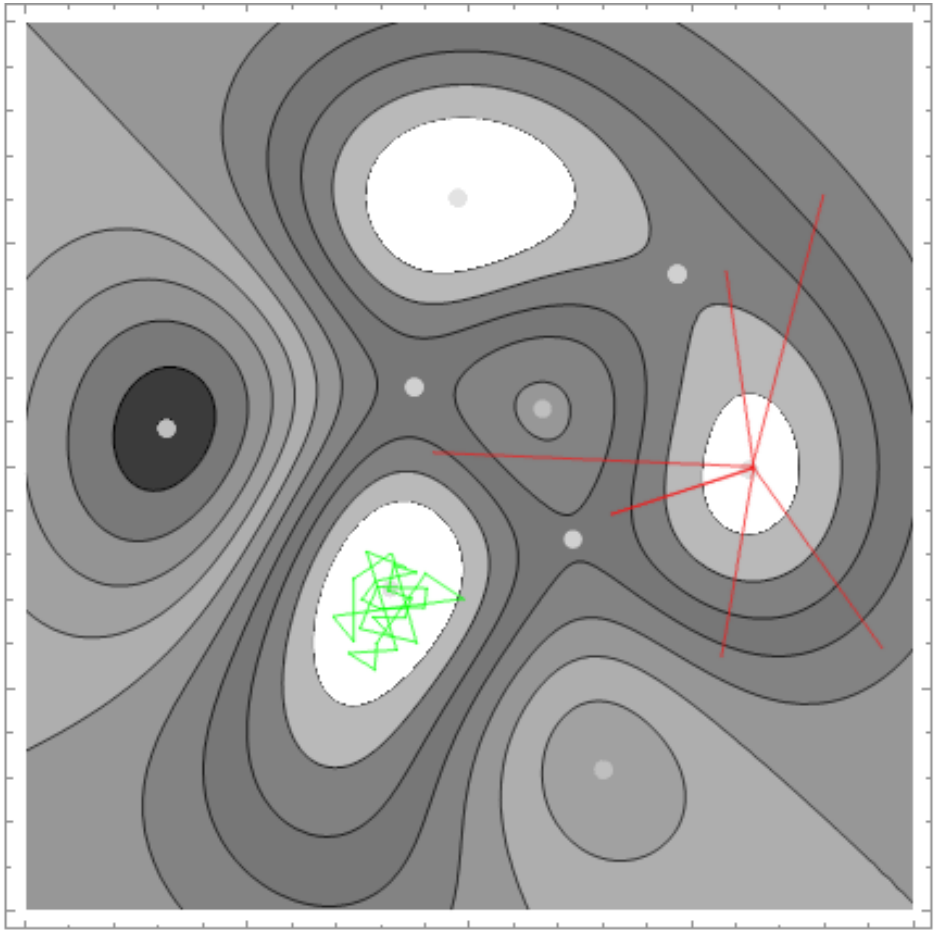


Figure: MCMC with either small (green) or large (red) proposals

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.

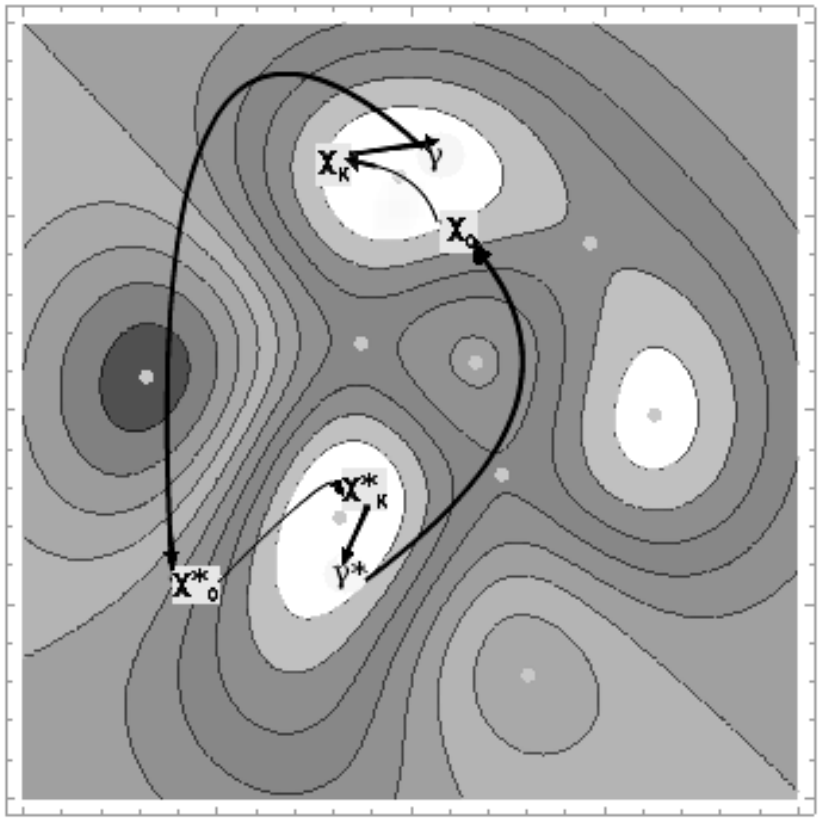


Figure: Locally optimized with randomization proposals

Application of MCMC with mode jumping proposals

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

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

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

Hence we also obtain alternative MCMC estimators of posterior marginal probabilities

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

- 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

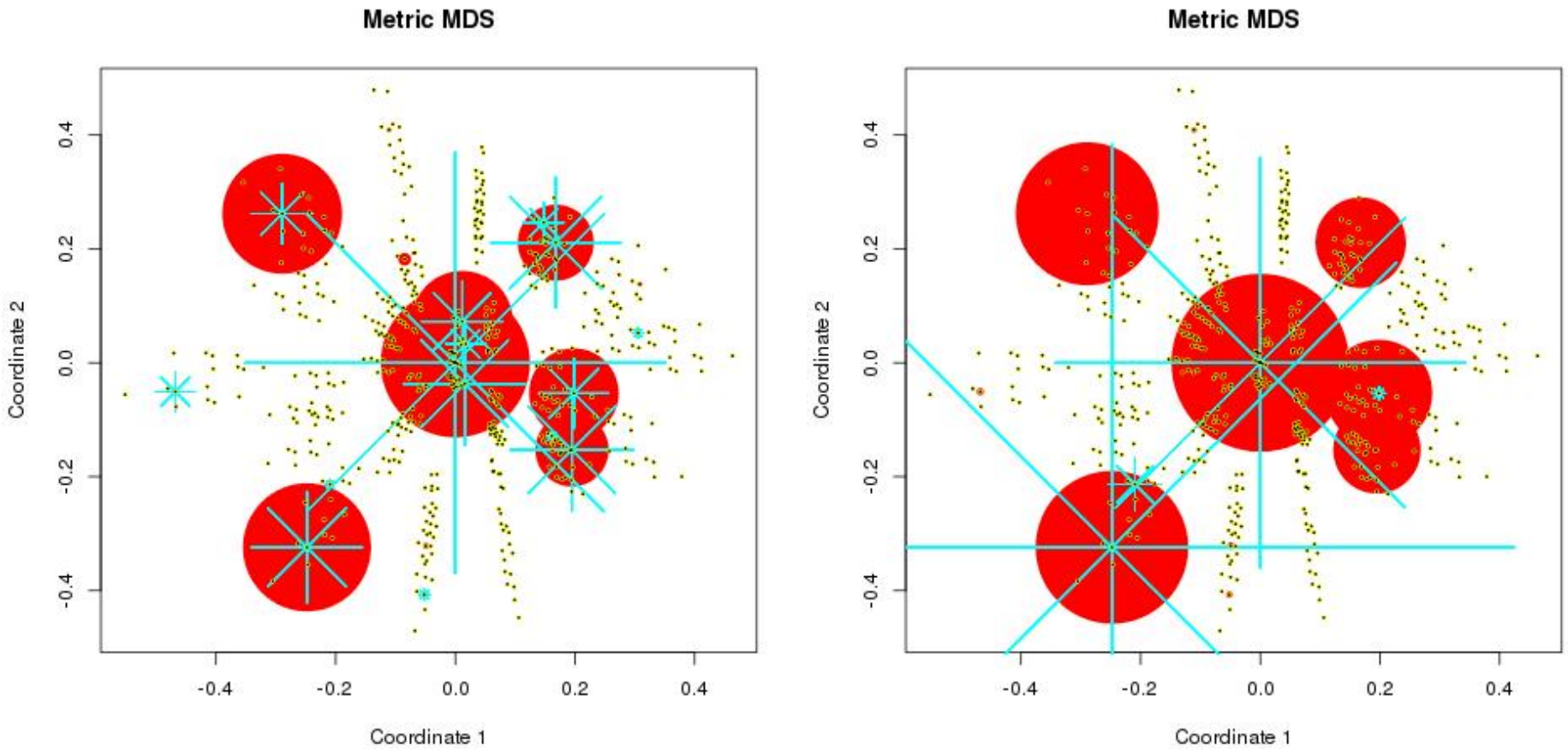


Figure: MDS plots with posterior modes of all found solutions for the approaches

The protein activity data. 2^{88} models. Multiple modes

Linear Bayesian regression with a Gelman's g-prior addressed:

$$y_t = y|m_t \sim N(m_t, \sigma_t) \tag{11}$$
$$m_t = \beta_0 + \sum_{i=1}^{p} \gamma_i \beta_i X_{t,i} \tag{12}$$
$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta = g(X'_\gamma X_\gamma)^{-1}), u = \sum_{i=1}^{p} \gamma_i \tag{13}$$
$$\sigma_t \sim N(0, \sigma_b) \tag{14}$$
$$\gamma_i \sim Binom(1, q = 0.5). \tag{15}$$

Analytical integration is possible. Marginal likelihoods become:

$$p(\mathbb{D}|\boldsymbol{\gamma}) \propto (1+g)^{(T-p-1)/2}(1+g[1-R^2_\gamma])^{-(T-1)/2}. \tag{16}$$

Here R^2_γ is the usual coefficient of determination of a linear regression model and g is the hyper-parameter of the g-prior. Notice that we set $g = T$.

The protein activity data. 2^{88} models. Multiple modes

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.

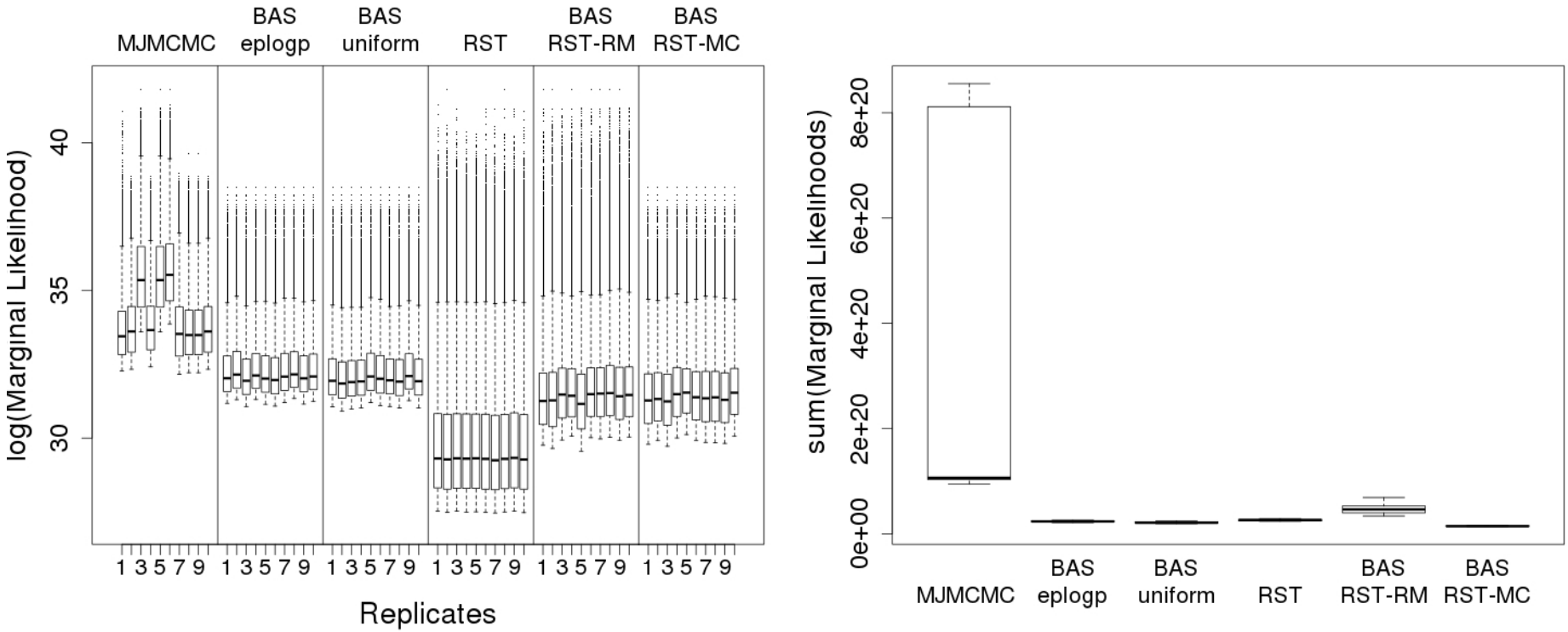


Figure: 100000 best mliks found (left) and posterior masses captured (right). Bayesian linear regression with a g-prior is addressed, since no other packages (to our awareness) manage model selection in GLMM

The protein activity data. 2^{88} models. Multiple modes

Checking convergence. Marginal inclusion probabilities

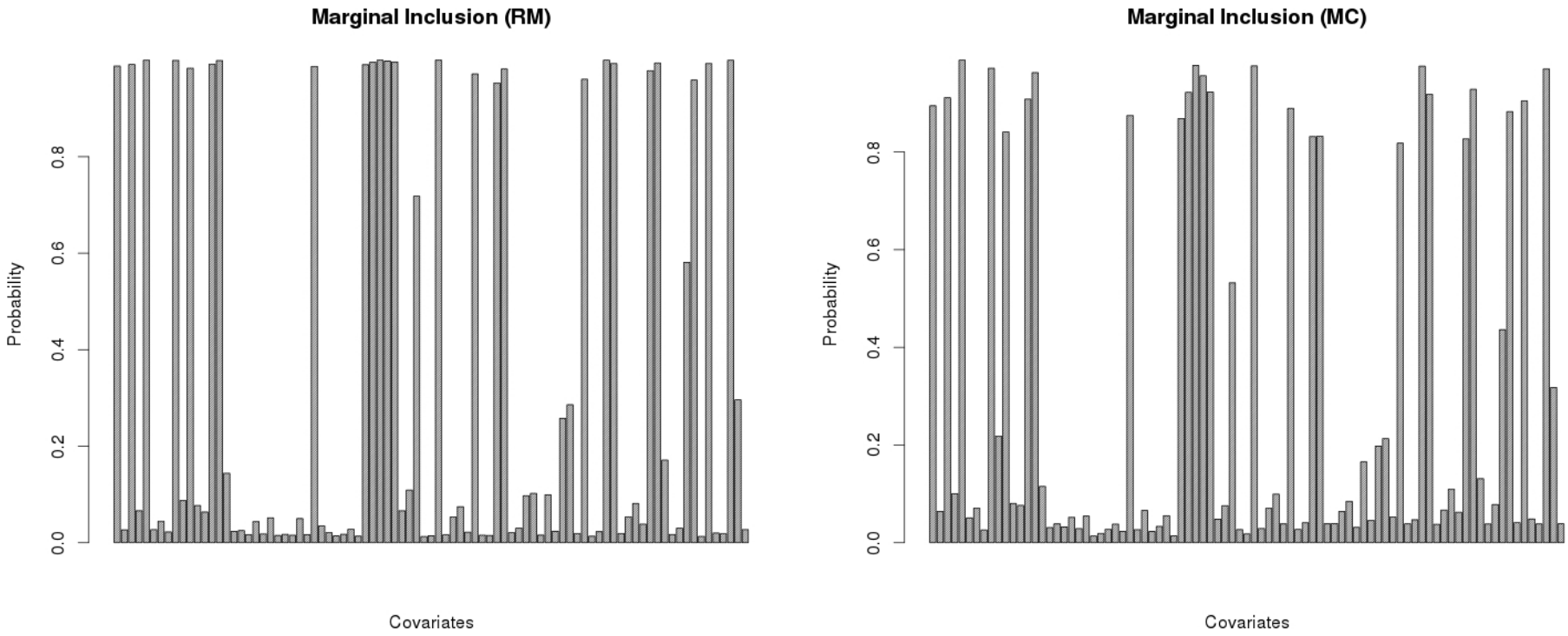


Figure: Comparison of marginal inclusion probabilities obtained by the Bayes formula and MCMC approximations from the best run of MJMCMC with $8.56e+20$ posterior mass captured

Bayesian model averaging

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

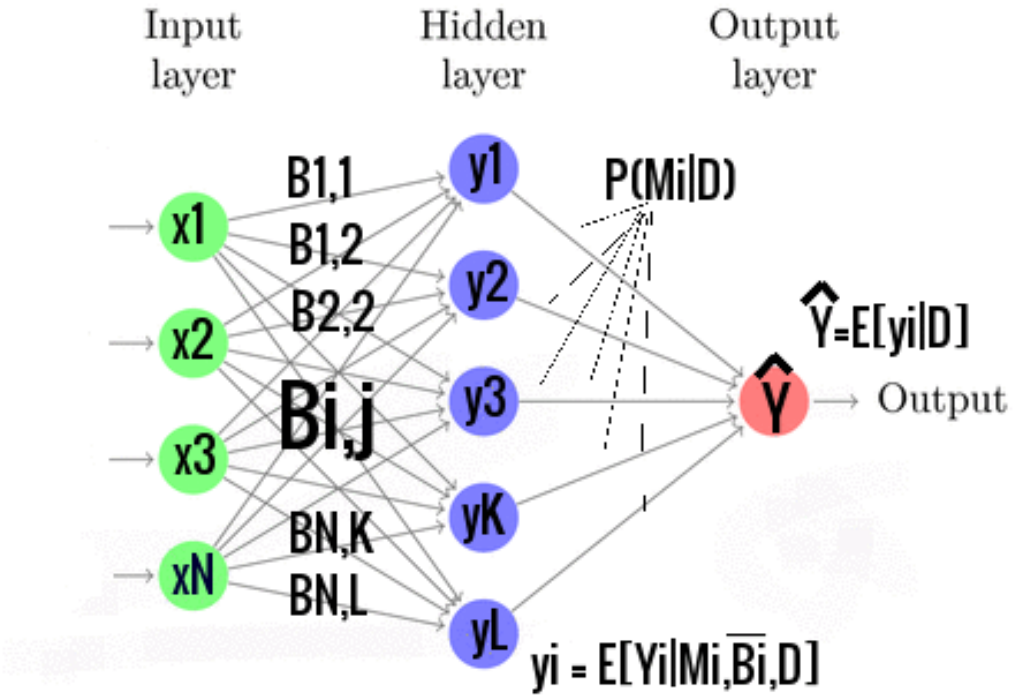
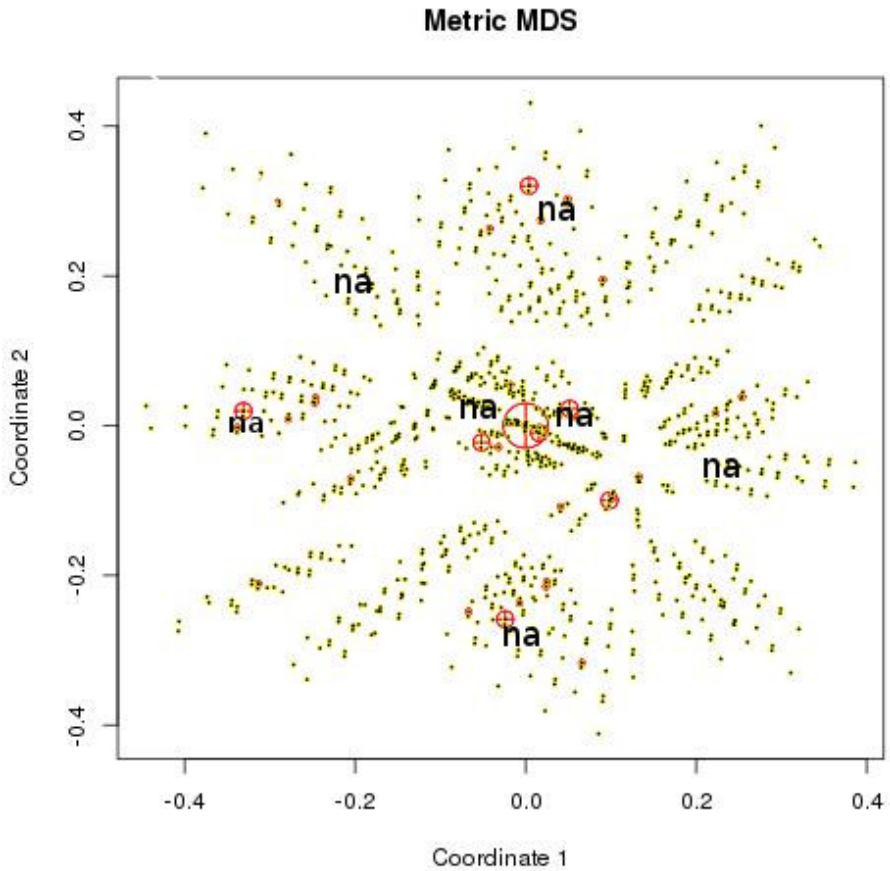


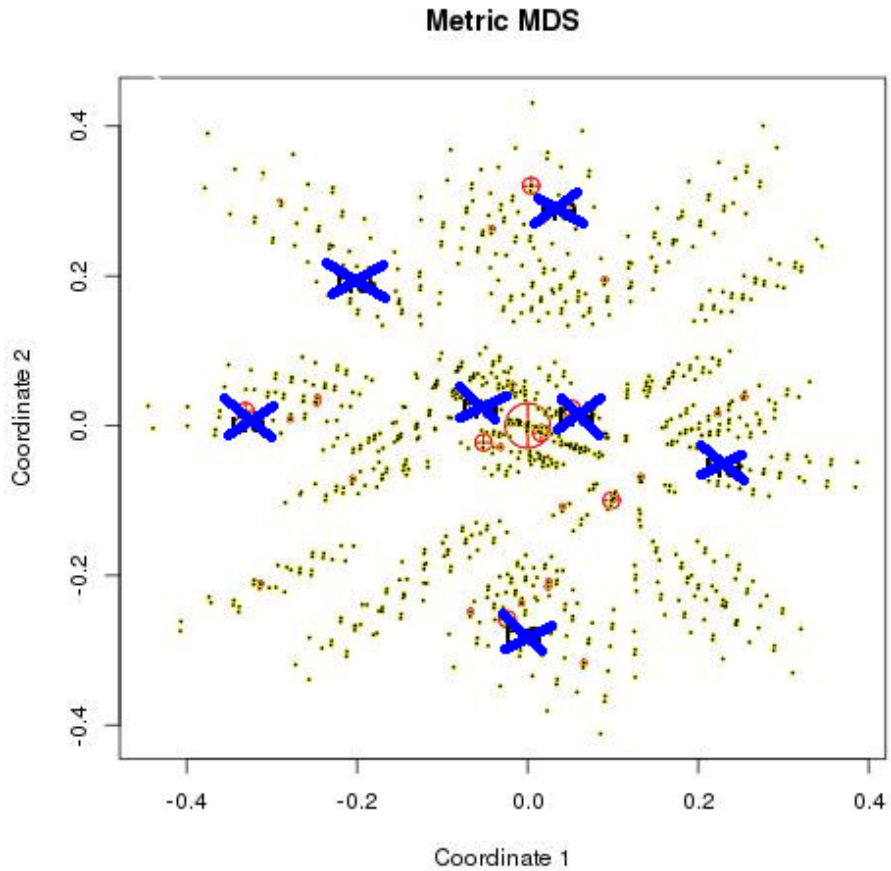
Figure: Bayesian model averaging

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

Missing data handling in predictions is easy and intuitive

- Delete models containing NA for the corresponding prediction from $\mathbb{V}$.
- Recalculate the posteriors.
- Get model averaged predictions.





Application. NEO objects classification. Problem

- **Observations**: Asteroid is a NEO (PHA) object or not (Phocaea)
- **Covariates**: 20 different covariates describing objects
- **Logistic Bayesian regression** addressed

$$y_t = y|p_t \sim Binom(1, p_t) \tag{17}$$

$$p_t = \frac{e^{\gamma_0\beta_0+\sum_{i=1}^{p}\gamma_i\beta_i x_{t,i}}}{1+e^{\gamma_0\beta_0+\sum_{i=1}^{p}\gamma_i\beta_i x_{t,i}}} \tag{18}$$

$$\beta|\gamma \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta = g(X'_\gamma X_\gamma)^{-1}), u = \sum_{i=1}^{p}\gamma_i \tag{19}$$

$$\gamma_i \sim Binom(1, q = 0.5). \tag{20}$$

Application. NEO objects classification. Results

$$\|\textbf{training set}\| = 64, \ \|\textbf{test set}\| = 20720, \ \|\textbf{missing data}\| = 10090$$

Subset	‖Hidden‖	Precision	FNR	FPR
$\mathbb{V}^0$	20005	99.95656%	0.05670945 %	0.01510117%
$\mathbb{V}^0$: 10912 rows with NA	20005	99.30502%	0.05670944 %	2.01272800%
$\mathbb{V}^1$	10090	99.95656%	0.05670945 %	0.01510117%
$\mathbb{V}^1$: 10912 rows with NA	10090	99.29054%	0.05670944 %	2.05621300%
$\mathbb{V}^2$	2512	99.80212%	0.05670945 %	0.49594239%
$\mathbb{V}^2$: 10912 rows with NA	2512	99.24228%	0.06379359 %	2.18643800%
$\mathbb{V}^3$	412	99.46429%	0.04253813 %	1.56110622%
$\mathbb{V}^3$: 10912 rows with NA	412	96.94015%	0.03545094 %	8.67586200%
$\mathbb{V}^4$	80	99.19402%	0.02836276%	2.40271201%
$\mathbb{V}^5$	4	90.00483%	0.04962427 %	23.7651171%
$\text{argmax}_{\boldsymbol{\gamma}\in\mathbb{V}^1}\{p_{\mathbb{V}}(\boldsymbol{\gamma}\|\mathbb{D})\}$	1	82.83301%	0.07087675 %	34.8839473%
Wake up NEO - no NA	?	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

Further (partly current) research

- Automatic creation of additional covariates based on the polynomes and interactions of the original ones as well as sigmoid functions of them (automatic feature extraction), based on an outer genetic algorithm (already implemented)
  I(erf(I(-(X37)((X54))))) added after 2^8 steps*
  I((I(-(X23)((X57))))) added after 2^{12} steps*
  ...
  I(tanh(I((X73)))) replaced I((I((X81)((X68))))) after 2^{16} steps*
  I((I((X71)((X69))))) replaced I(erf(I(((X54))))) after 2^{18} steps*
  ...
- Allowing the search over different choices of the random effect structures (to be addressed)
- Allowing the search over different choices of the response distributions (to be addressed)

Concluding remarks

- We introduced the MJMCMC algorithm
  - estimating posterior model probabilities
  - Bayesian model averaging and selection
- *EMJMCMC* R-package is developed
  - `http://aliaksah.github.io/EMJMCMC2016/`
  - flexibility in the choice of methods
    - marginal likelihoods
    - model selection criteria
  - extensive parallel computing is available
  - vectorized predictions with NA hadling is incorporated
- Results showed that MJMCMC
  - performs well in terms of the search quality
  - addresses a more general class of models than competitors
  - provides nice predictive performance in the applications

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.