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

Imaguru business club, Minsk

06.09.2016

Introduction. Issues

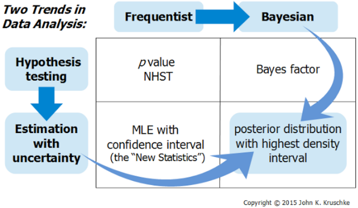
- GLMM are used in a wide range of different applications for
  - Inference
  - Predicton
- More sources of data → more hypothetical explanatory variables →
  - Model selection
  - Model averaging
- Posterior marginal model probabilities are used to
  - Estimate quality of the models
  - Serve as weights 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

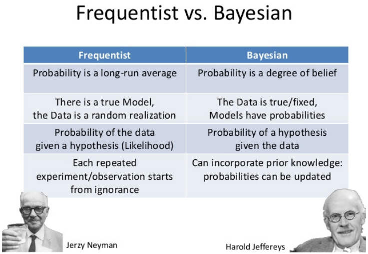
Bayesian vs. Frequentist statistics

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

vs.

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





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

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`)

A reminder example

Logistic Bayesian regression

$$y_t = y|p_t \sim Binom(1, p_t) \quad (1)$$

$$p_t = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_{t,i}}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_{t,i}}} \quad (2)$$

$$\boldsymbol{\beta}|\gamma \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta = g(X'_\gamma X_\gamma)^{-1}), u = \sum_{i=1}^{p} \gamma_i \quad (3)$$

Notice that priors appear here

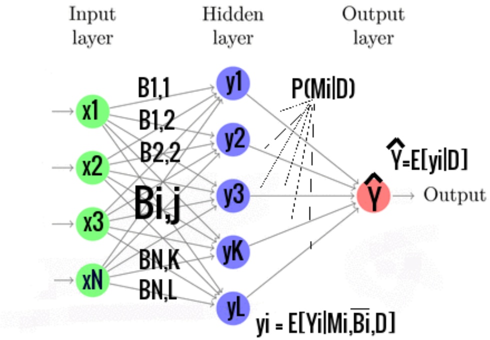


Figure: Bayesian Artificial Neuron Network for Classification

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

Bayesian Generalized Linear Mixed Model

$$
\begin{align}
Y_t|\mu_t \sim \mathrm{f}(y|\mu_t), t \in \{1, ..., T\} \tag{4}\\
\mu_t = g^{-1}(\eta_t) \tag{5}\\
\eta_t = \gamma_0\beta_0 + \textstyle\sum_{i=1}^{p} \gamma_i\beta_i X_{ti} + \delta_t \tag{6}\\
\boldsymbol{\delta} = (\delta_1, ..., \delta_T) \sim N_T(\mathbf{0}, \boldsymbol{\Sigma}_b). \tag{7}
\end{align}
$$

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

$$\gamma_i \sim Binom(1, q) \tag{8}$$
$$q \sim Beta(\alpha_q, \beta_q) \tag{9}$$
$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), u = \sum_{i=1}^{p} \gamma_i \tag{10}$$
$$\boldsymbol{\psi} \sim \varphi(\boldsymbol{\psi}). \tag{11}$$

- 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 $\mathfrak{S}$

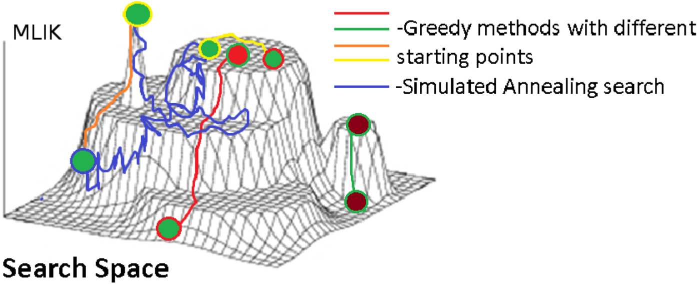
Procedure

- **Notice that** $p(\gamma, \theta|\mathbb{D}) = p(\theta|\gamma, \mathbb{D})p(\gamma|\mathbb{D})$
- $p(\theta|\gamma, \mathbb{D})$ and $\log p(\mathbb{D}|\gamma)$ can be efficiently obtained by INLA
- **Notice that** $p(\gamma|\mathbb{D}) = \frac{e^{\log p(\mathbb{D}|\gamma)+\log p(\gamma)}}{\sum_{\gamma' \in \Omega_\gamma} e^{\log p(\mathbb{D}|\gamma')+\log p(\gamma')}}$
- $\widehat{p}(\gamma|\mathbb{D}) = \frac{e^{\log p(\mathbb{D}|\gamma)+\log p(\gamma)}}{\sum_{\gamma' \in \mathbb{V}} e^{\log p(\mathbb{D}|\gamma')+\log p(\gamma')}}$
- $\mathbb{V}$ is the **subspace** of Ω_γ to be **efficiently explored**
- **Notice 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 $\Longrightarrow$
- **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

Possible ways to explore $\mathbb{V} \subset \Omega_\gamma$

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
- MCMC with mode jumping proposals seems to be a good idea



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.

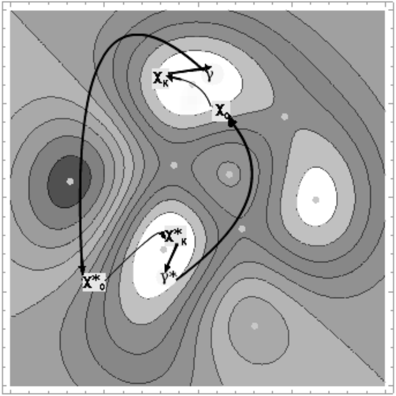
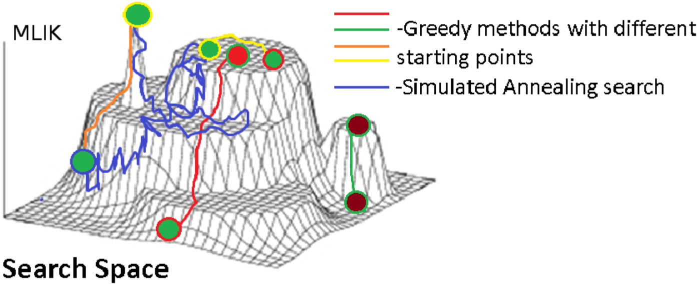


Figure: Locally optimized with randomization proposals

Local combinatorial optimizers

- **Greedily optimized local improvements** (implemented)
- *Simulated annealing based local improvements* (implemented)
- *MCMC based local improvements* (implemented)
- Other local metaheuristics (TA, ant colony optimization, local genetic algorithms, etc) (not addressed yet)
- *Combinations of them* (implemented)



Allowed transitions

Proposal $q(\gamma^*\|\gamma)$	Label
$\frac{\prod_{i\in\{j(1),\dots,j(S)\}}\rho_i}{\binom{p}{S}(\eta-\zeta+1)}$	*Random change with random size of the neighborhood*
$\frac{\prod_{i\in\{j(1),\dots,j(S)\}}\rho_i}{\binom{p}{S}}$	*Random change with fixed size of the neighborhood*
$\frac{1}{\binom{p}{S}(\eta-\zeta+1)}$	*Swap with random size of the neighborhood*
$\binom{p}{S}^{-1}$	*Swap with fixed size of the neighborhood*
$\frac{1-\mathbb{I}\left(\sum_i^p(\gamma_i)=P\right)}{P-\sum_i^p\gamma_i+\mathbb{I}\left(\sum_i^p(\gamma_i)=P\right)}$	*Uniform addition of a covariate*
$\frac{1-\mathbb{I}\left(\sum_i^p(\gamma_i)=0\right)}{\sum_i^p\gamma_i+\mathbb{I}\left(\sum_i^p(\gamma_i)=P\right)}$	*Uniform deletion of a covariate*

Table: Types of proposals suggested for the moves between the models during MCMC procedure. Here S is either deterministic or random $S \sim Unif\{\zeta,...,\eta\}$ size of the neighborhood; $\rho_i, i\in\{j(1),...,j(S)\}$ is the probability of inclusion of variable $\gamma_i, i\in\{j(1),...,j(S)\}$, which can be either deterministic or addaptive when $\rho_i, i\in\{j(1),...,j(S)\}$ are adaptively updated approximations of the marginal inclusion probabilities; $\mathbb{I}(\cdot)$ is the identity function; p is the total number of covariates.

Application of MCMC with mode jumping proposals

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

$$r_m(\gamma_j, \gamma_k) = \min\left\{1, \frac{p(\mathbb{D}|\gamma_k)p(\gamma_k)\mathrm{q_s}(\gamma_j|\gamma_{j_{K-1}})}{p(\mathbb{D}|\gamma_j)p(\gamma_j)\mathrm{q_s}(\gamma_k|\gamma_{k_{K-1}})}\right\}. \tag{12}$$

- $\mathrm{q_s}(.|.)$ is the kernel of randomization at the end.

Hence we also obtain alternative MCMC estimators of posterior marginal probabilities

$$\tilde{p}(\gamma|\mathbb{D}) = \frac{\sum_{i=1}^{W} \mathbb{I}(\gamma_i = \gamma)}{W} \xrightarrow[W\to\infty]{d} p(\gamma|\mathbb{D}). \tag{13}$$

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

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

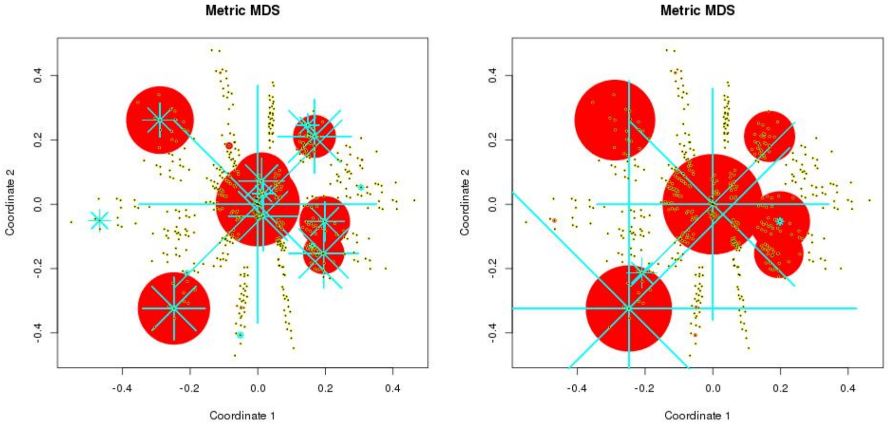


Figure: MDS plots with posterior modes of all found solutions for the approaches

Application to 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

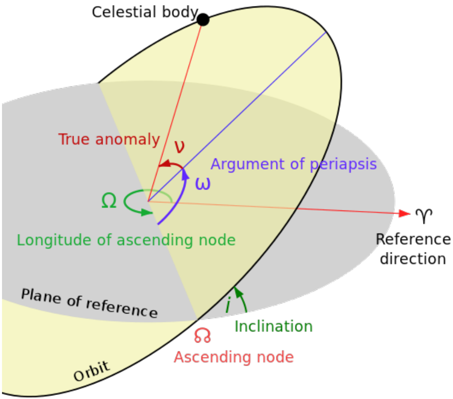
$$y_t = y|p_t \sim Binom(1, p_t) \quad (14)$$

$$p_t = \frac{e^{\gamma_0\beta_0+\sum_{i=1}^{p}\gamma_i\beta_i X_{t,i}}}{1+e^{\gamma_0\beta_0+\sum_{i=1}^{p}\gamma_i\beta_i X_{t,i}}} \quad (15)$$

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_u(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta = g(X_\gamma' X_\gamma)^{-1}), u = \sum_{i=1}^{p}\gamma_i \quad (16)$$

$$\gamma_i \sim Binom(1, q = 0.5). \quad (17)$$

NEO objects classification. NASA space challenge



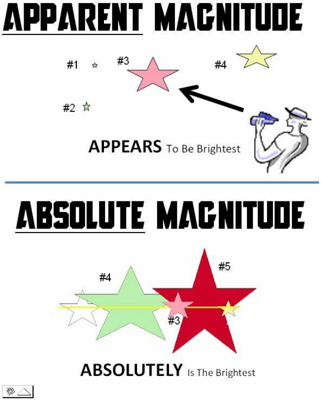


Figure: Orbital elements(left) by Lasunncty (talk), CC BY-SA 3.0 and absolute vs apparent magnitude (right) by Mrscreath(`http://mrscreath.blogspot.com`)

NEO objects classification. NASA space challenge

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$\times$Semi major axis; Mean anomaly$\times$Semi major axis^2 $\in \mathbf{R}^+$; Mean anomaly$\times$Semi major axis^3 $\in \mathbf{R}^+$; Argument of perihelion$\times$Semi major axis $\in \mathbf{R}^+$; Argument of perihelion$\times$Semi major axis^2 $\in \mathbf{R}^+$; Argument of perihelion$\times$Semi major axis^3 $\in \mathbf{R}^+$; Longitude of the ascending node$\times$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).

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

NEO objects classification. Inference

Posterior inclusion probabilities and posterior model probabilities

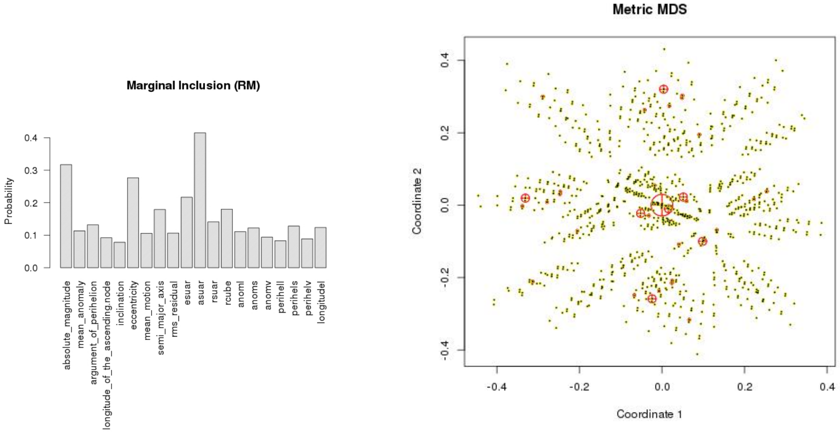


Figure: Comparison of marginal inclusion probabilities of the covariates (left) and models on the whole (right)

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!!!

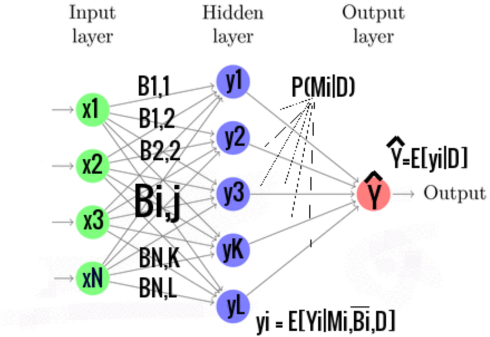
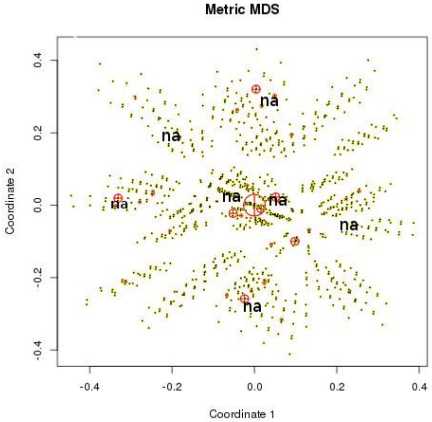


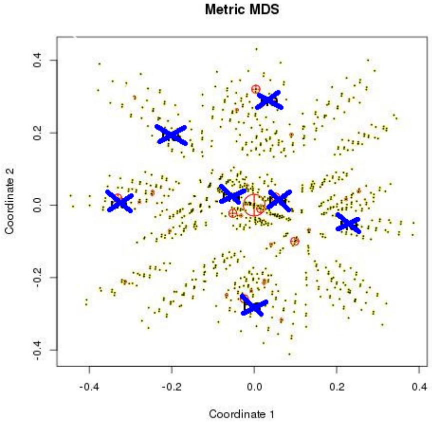
Figure: Bayesian Artificial Neuron Network for Classification

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

NEO objects classification. Missing data handling

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





NEO objects classification. Results

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

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%
$\text{argmax}_{\gamma\in\mathbb{V}^1}\{p_{\mathbb{V}}(\gamma\|\mathbb{D})\}$	1	82.83301%	0.07087675 %	34.8839473%
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

Multicore and shared memory issues

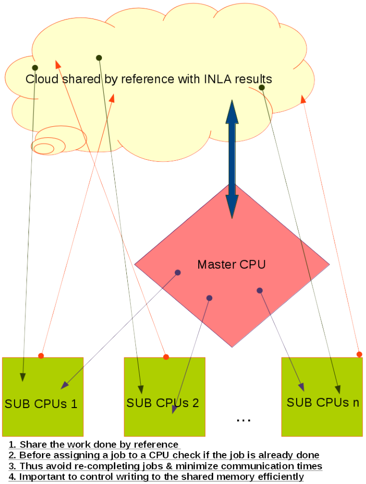


Figure: Multiprocessing architecture

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.

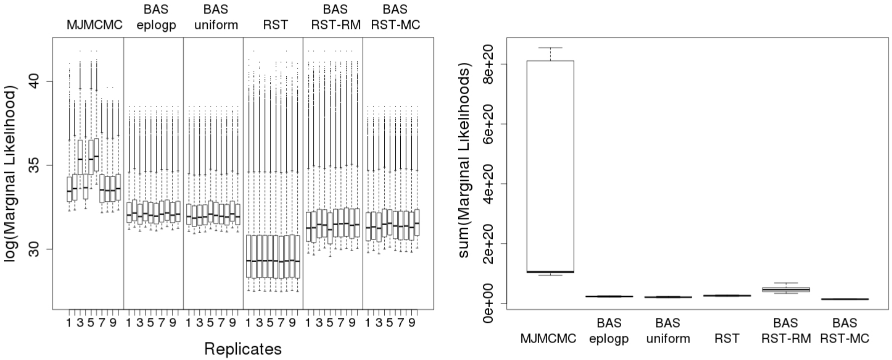


Figure: 100000 best mliks found (left) and posterior masses captured (right). Bayesian linear regression with a g-prior is addressed, since no other packages (to our awareness) manage model selection in GLMM

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

Checking convergence. Marginal inclusion probabilities

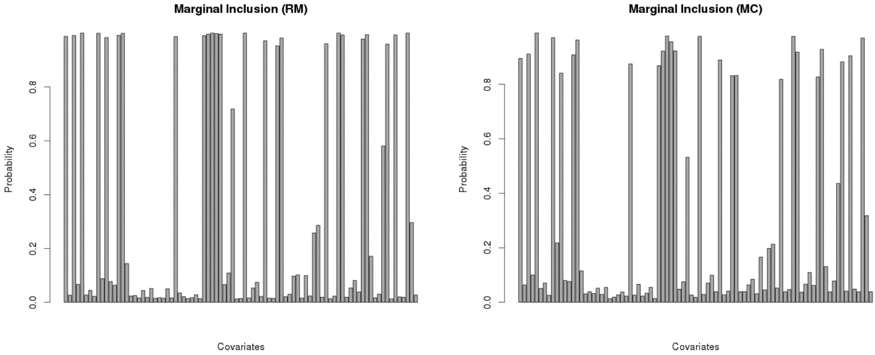


Figure: Comparison of marginal inclusion probabilities obtained by the Bayes formula and MCMC approximations from the best run of MJMCMC with $8.56e+20$ posterior mass captured

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 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.