# Variable selection in binomial regression with latent Gaussian field models for analysis of epigenetic data

Aliaksandr Hubin

University of Oslo

*aliaksah@math.uio.no*

19.11.2015

# Overview

# Introduction

- More precise estimation of the methylation status of locations
- Discovery of methylated and unmethylated regions and corresponding local and global structures:
  - Represented by nucleotides sequences patterns (CG-islands, CPG-islands)
  - Represented by such structures as genes on the whole, promoters, expressions and their sequences
- Finding covariates (location within the gene, underlying genetic structure, chromosome etc.) significantly influencing methylation patterns along the genome
- Linking genetic and epigenetic data to phenotypic responses (expressions of genes, presence of transposons, etc.) in a statistically significant way

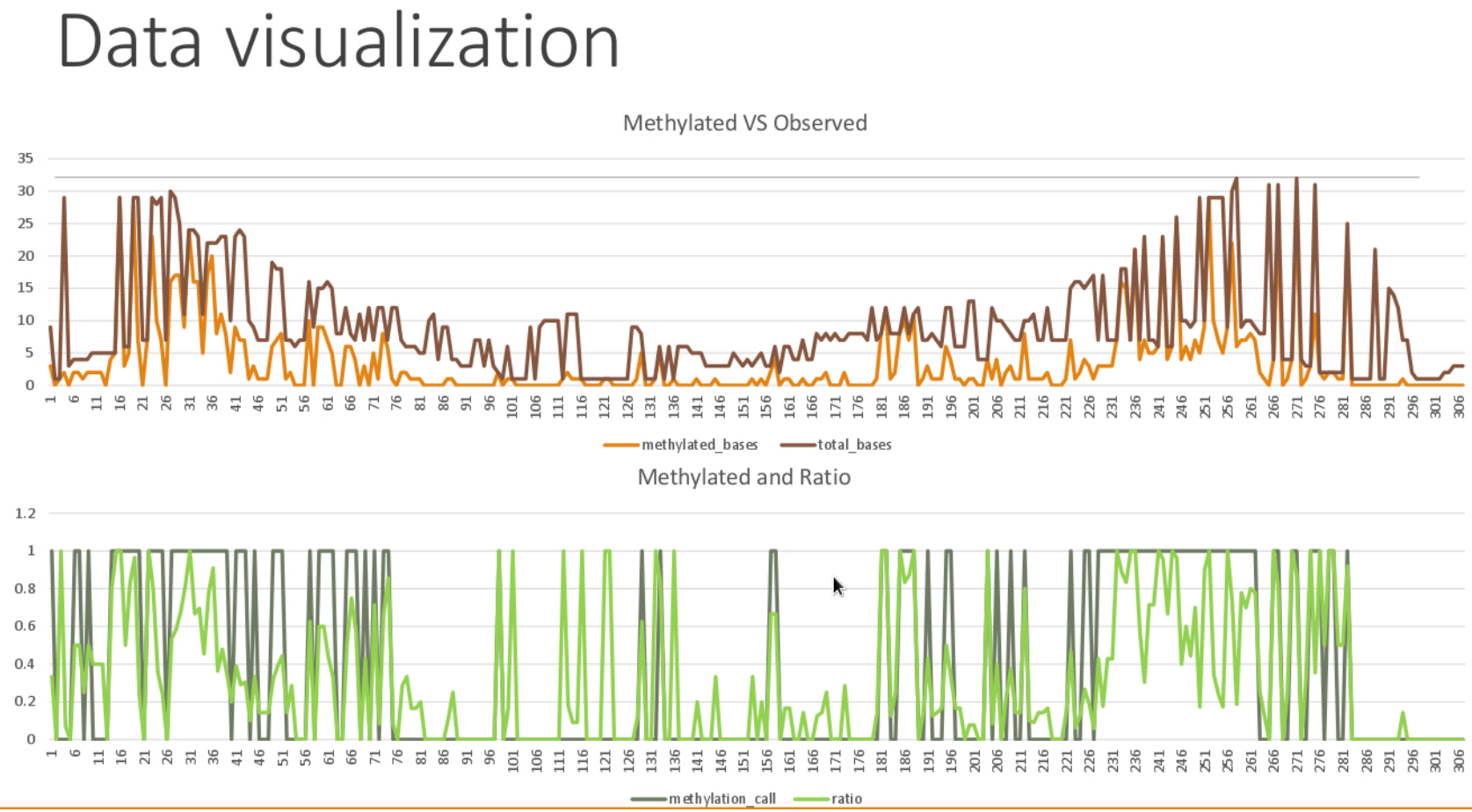



Figure: Total reads and methylated reads for some part of the genome

# The model

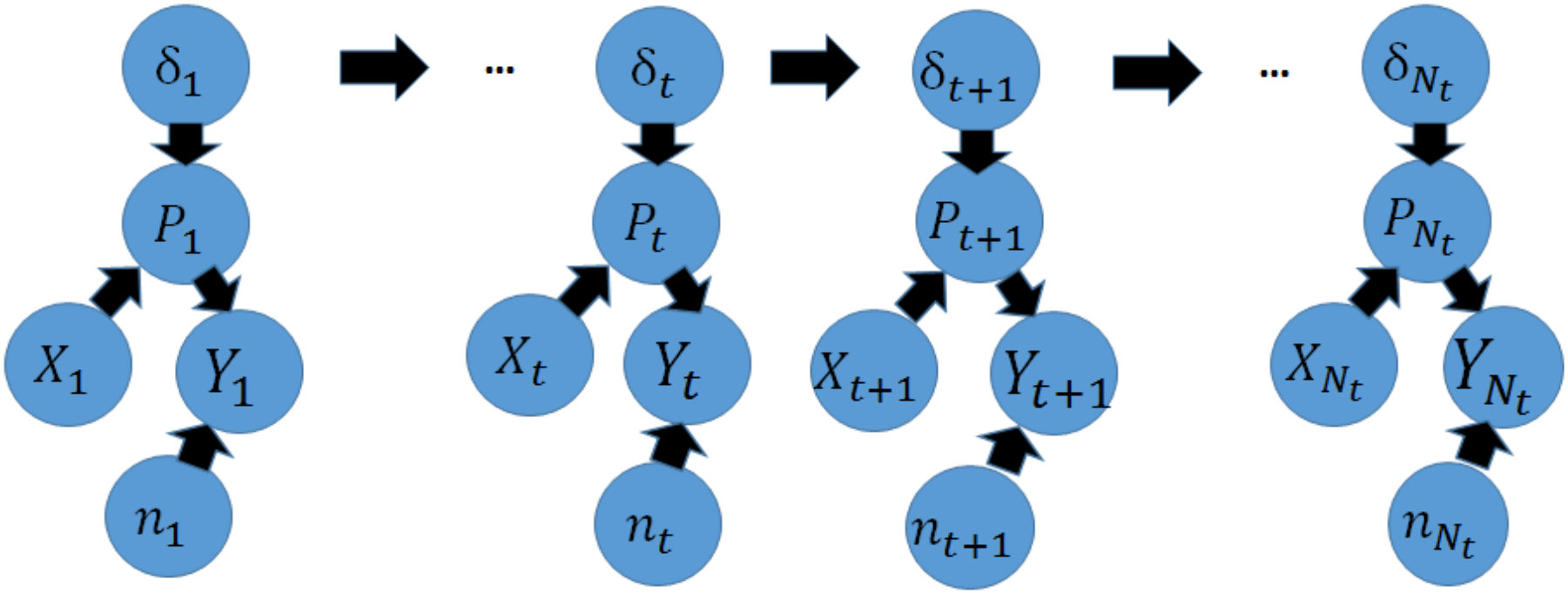



Figure: The model

# The model

Generalized Binomial Regression With Gaussian Latent Field Model **(GBR-WGLF)**

$$\Pr(y_t = y | n_t = n, p_t) = \binom{n}{y} p_t^y (1 - p_t)^{n-y} \tag{1}$$

$$p_t = \frac{e^{\gamma_0 \beta_0 + \sum_{i=1}^{M} \gamma_i \beta_i X_{t,i} + \delta_t}}{1 + e^{\gamma_0 \beta_0 + \sum_{i=1}^{M} \gamma_i \beta_i X_{t,i} + \delta_t}} \tag{2}$$

$$\delta_t = \rho \delta_{t-1} + \epsilon_t \tag{3}$$

- $y_t \in \{1, ..., N_t\}$ is the number of methylated reads per loci $t$
- $n_t \in \mathbb{N}$ is the total number of reads per loci $t$
- $\beta_i \in \mathbb{R}, i \in \{0, ..., N_\gamma\}$ are regression coefficients of the covariates of the model
- $\gamma_i \in \{0, 1\}, i \in \{0, ..., N_\gamma\}$ are latent indicators, defining if covariate $i$ is included into the model
- $\rho \in \mathbb{R}$ is the coefficient of latent $AR(1)$ process

# Hyper-parameters of the model

$$\gamma_i \sim Binom(1, q) \tag{4}$$
$$q \sim Beta(\alpha_q, \beta_q) \tag{5}$$
$$\beta_i|\gamma_i \sim \mathbb{I}(\gamma_i = 1)N(\mu_\beta, \sigma^2_\beta) \tag{6}$$
$$\begin{pmatrix}\psi_1 \\ \psi_2\end{pmatrix} \sim N_2(\mu_{\rho,\epsilon}, \Sigma_{\rho,\epsilon}) \tag{7}$$
$$\epsilon_t \sim N(0, \sigma^2_\epsilon) \tag{8}$$

- $\psi_1 = \log \frac{1}{\sigma^2_{\epsilon,t}}(1-\rho^2)$ and $\psi_2 = \log \frac{1+\rho}{1-\rho}$ are scaled hyper-parameters of the latent model
- $\epsilon_t$ is the error term of $AR(1)$ model
- $q$ is the prior probability of including any covariate into the model

# Inference on the model

**Let:**

- $\mathbb{M} : \vec{\gamma}$ define a subset of parameters of GBRWGLF
- $\Theta = \{\vec{\beta}, \rho, \sigma_\epsilon^2\}$ define all other parameters of GBRWGLF
- $\Theta|\mathbb{M}$ define parameters of GBRWGLF model conditioned on fixed $\vec{\gamma}$, further addressed as a Binomial Regression With Gaussian Latent Field Model **(BRWGLF)** or simply a model
- $\exists L = 2^{N_\gamma + 1}$ different BRWGLF models

**Goals:**

- $\Pr(\mathbb{M}, \Theta|\mathbb{D})$ posterior distribution of parameters of GBRWGLF
- $\Pr(\mathbb{M}|\mathbb{D})$ marginal posterior distribution of the models
- Set of estimated BRWGLF performing well in terms of some model selection criteria (MAP, WAIC, DIC, MLIK) to explain the phenomena optimally

# Procedure

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

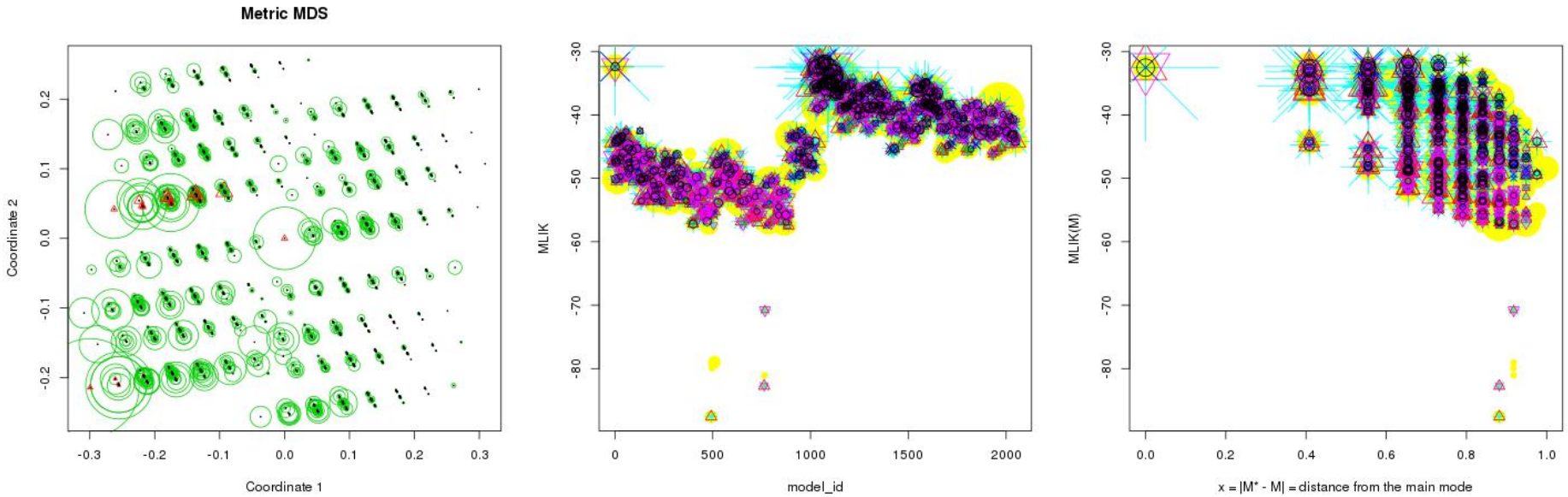
# Main challenges

Once again:

- Note that $\Pr(\mathbb{M}, \Theta|\mathbb{D}) = \Pr(\Theta|\mathbb{M}, \mathbb{D}) \Pr(\mathbb{M}|\mathbb{D})$
- $\Pr(\Theta|\mathbb{M}, \mathbb{D})$ and $\log \Pr(\mathbb{D}|\mathbb{M} = M)$ are obtained by INLA

Proceed with efficient exploration of $\mathbb{V}$ in the subspace of $\Omega_{\mathbb{M}}$ to estimate $\Pr(\mathbb{M} = M|\mathbb{D})$, $\underset{M \in \Omega_{\mathbb{M}}}{\operatorname{argmax}} \Pr(\mathbb{M} = M|\mathbb{D})$, and $\underset{M \in \Omega_{\mathbb{M}}}{\operatorname{argmax}}$ WAIC$(M)$

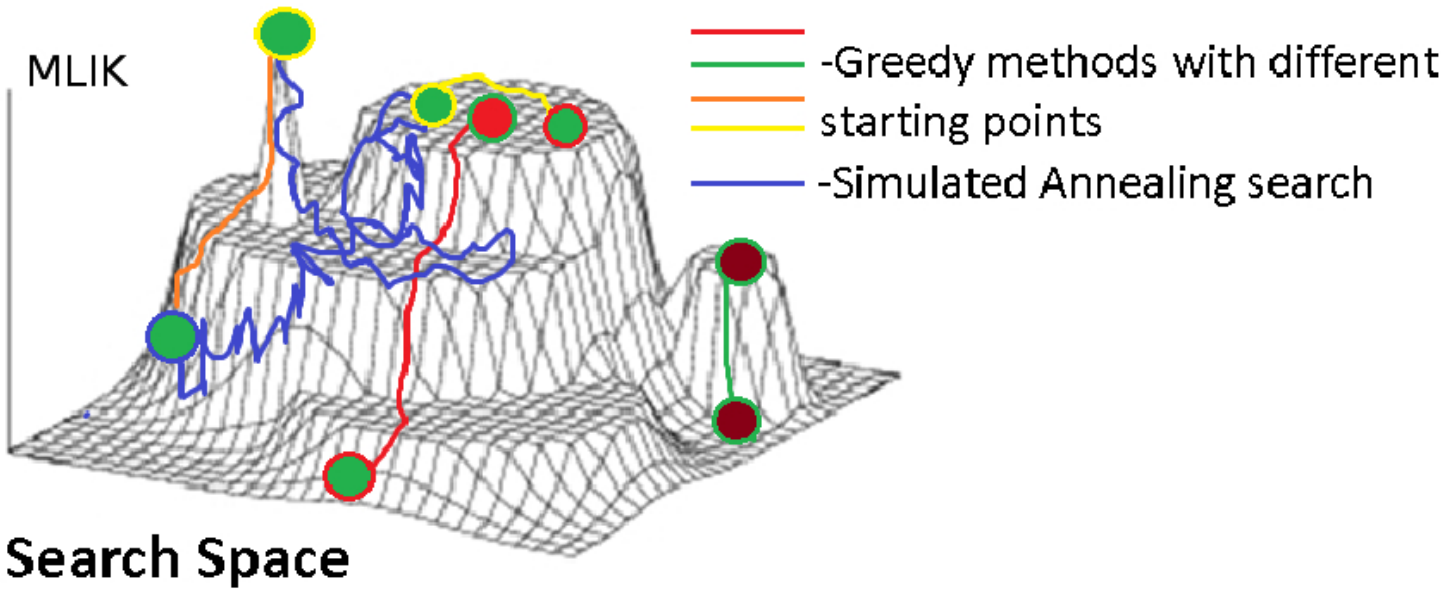
Main challenges are multimodality in $\Omega_{\mathbb{M}}$ and its size:



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

**Main challenges are multimodality in $\Omega_{\mathbb{M}}$ and its size.**

- Full enumeration of $\Omega_{\mathbb{M}}$ - infeasible for large dimensions
- Random walk in $\Omega_{\mathbb{M}}$ including simple MCMC - does not take advantage of the structure of $\Omega_{\mathbb{M}} \Longrightarrow$ too slow
- Greedy optimization - end up in local optima
- SA - ends up with random descent with almost no chance to change the mode
- Random walk with mode jumping proposals seems to be a good idea



# MCMC with mode jumping proposals

## Notice that

We address **Locally annealed**, **locally optimized**, and **locally multiple try simulated proposals** and their combination in a way to satisfy the detailed balance equation.

## Also notice

Also note that within this setting of locally optimized MTMCMC we get an alternative MCMC based approximations for posterior probabilities of the models, namely $\tilde{\Pr}(\mathbb{M} = M|\mathbb{D}) = \frac{\sum_{i=1}^{W} \mathbb{I}(M_i = M)}{W} \xrightarrow[W\to\infty]{d} \Pr(\mathbb{M} = M|\mathbb{D})$ and $\underset{i\in 1,\ldots,W}{\operatorname{argmax}} \text{WAIC}(M_i) \xrightarrow[W\to\infty]{} \underset{M \in \Omega_{\mathbb{M}}}{\operatorname{argmax}} \text{WAIC}(\mathbb{M} = M)$. Whist $\mathbb{V} = \bigcup_{i=1}^{W} M_i \xrightarrow[W\to\infty]{} \Omega_{\mathbb{M}}$. This allows us to verify the results.

# MTMCMC with SA proposals

We will now address SA based proposals

**Algorithm 1** Simulated annealing optimization

**procedure** ANNEAL($T_c, \Pr_a(.,.|.), f(.), x_0$) ▷ cooling schedule, acceptance probabilities, objective function and initial point
  $x \leftarrow x_0$
  $x_b \leftarrow x_0$
  **for** $t$ in $T_c$ **do** ▷ for all temperatures in the cooling schedule
    $x_c \leftarrow \mathbb{N}(x)$ ▷ pick a random neighbor of the current solution
    **if** $\mathsf{G}(x_c) > \mathsf{G}(x_b)$ **then**
      $x_b \leftarrow x_c$ ▷ update the best found solution
    **end if**
    **if** $\Pr_a(x, x_c|t) > u \sim Unif[0;1]$ **then**
      $x \leftarrow x_c$ ▷ accept the move with some probability
    **end if**
  **end for**
  **return** $x, x_b$ ▷ return the final solution
**end procedure**

# MTMCMC with SA proposals

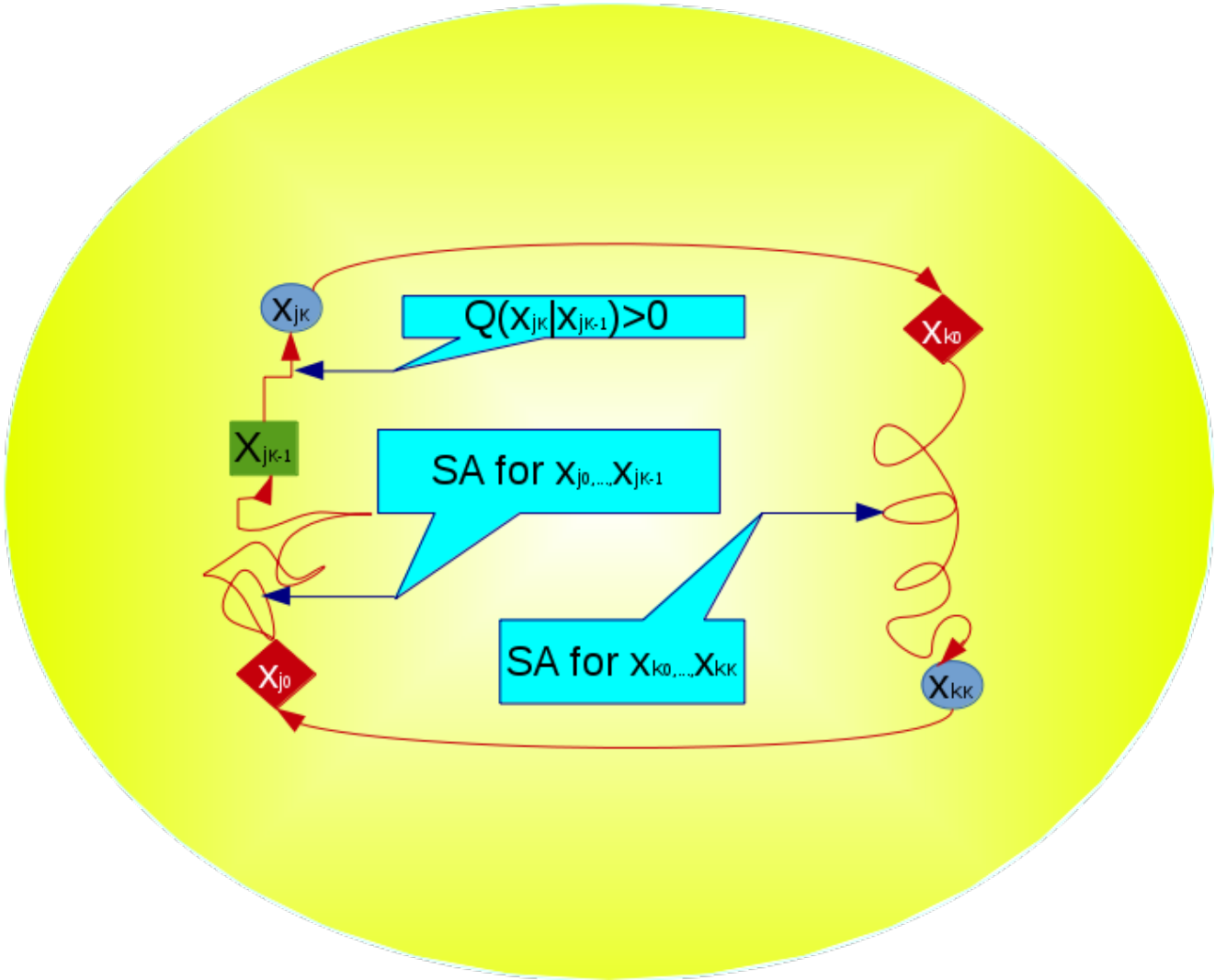



Figure: Simulated annealed symmetric proposals

# Multicore and shared memory issues

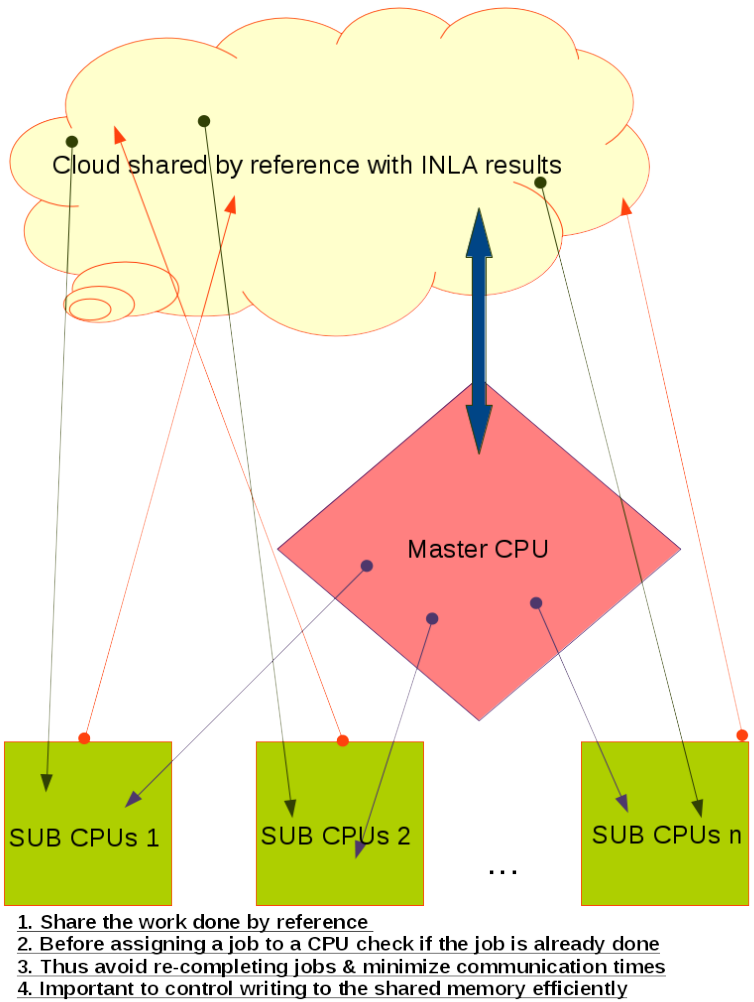



Figure: Multiprocessing architecture

# Results

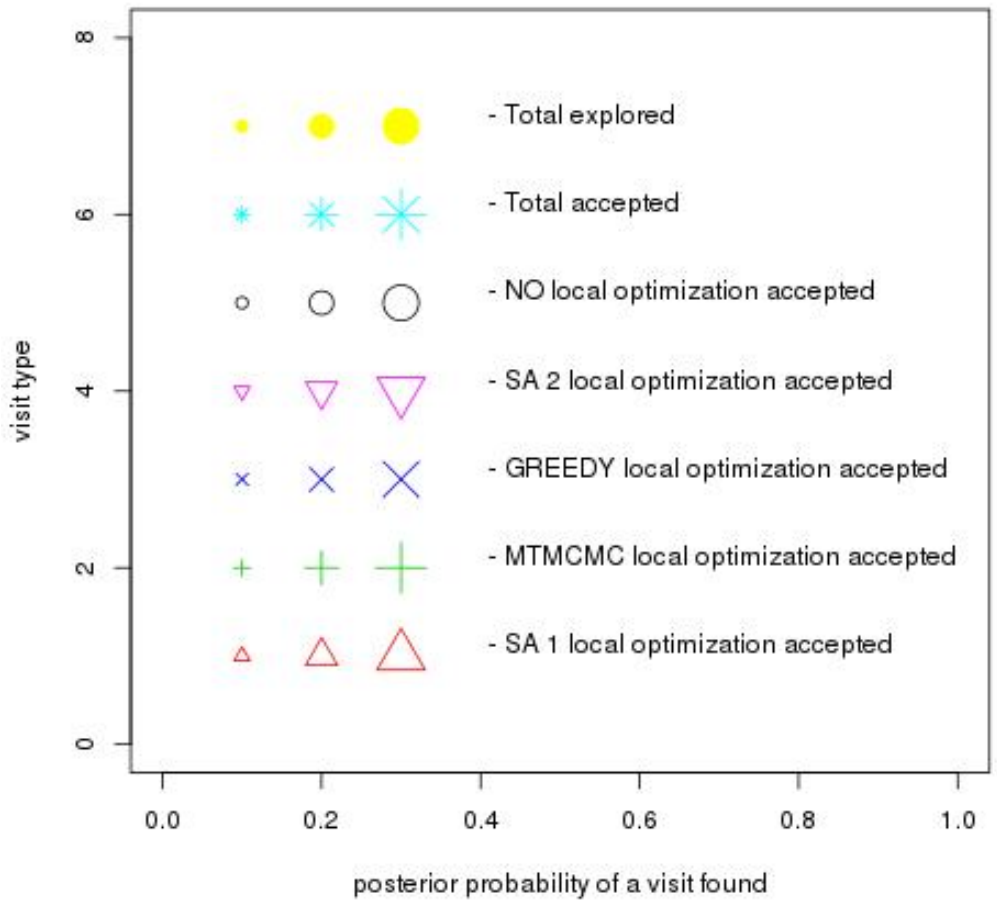
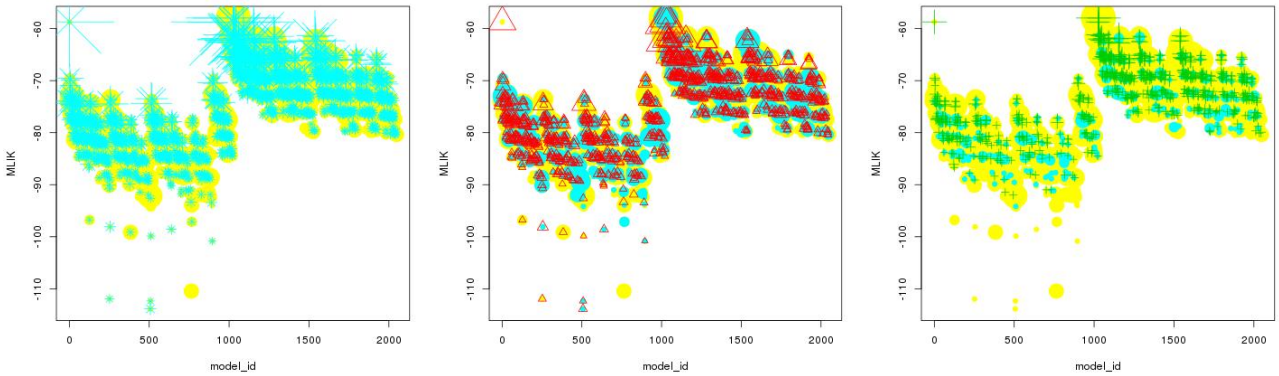
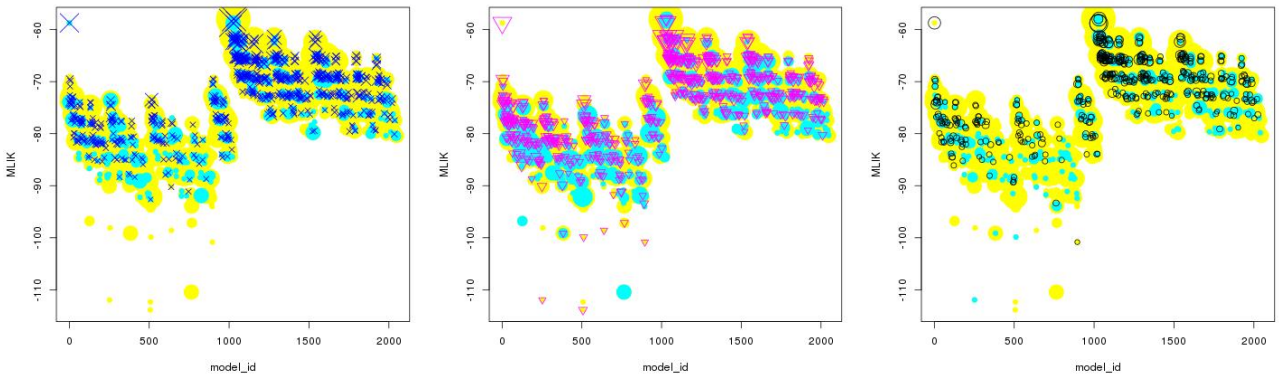



Figure: Notation

# Results




(a) totally explored (b) plus SA 1 (c) plus locMCMC




(d) plus greedy (e) plus SA 2 (f) no local opt.

Figure: MLIK against model index different methods

# Results

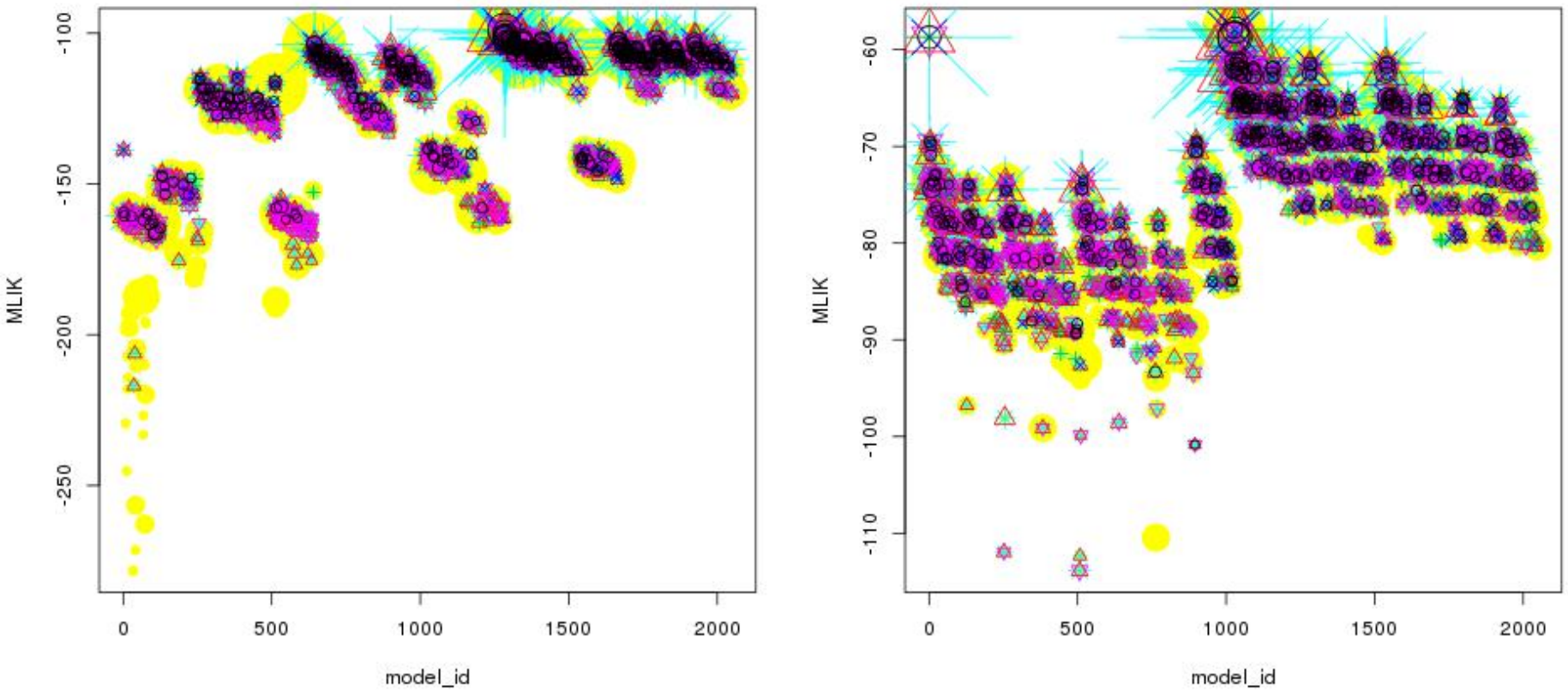



Figure: MLIK against model index

# Results

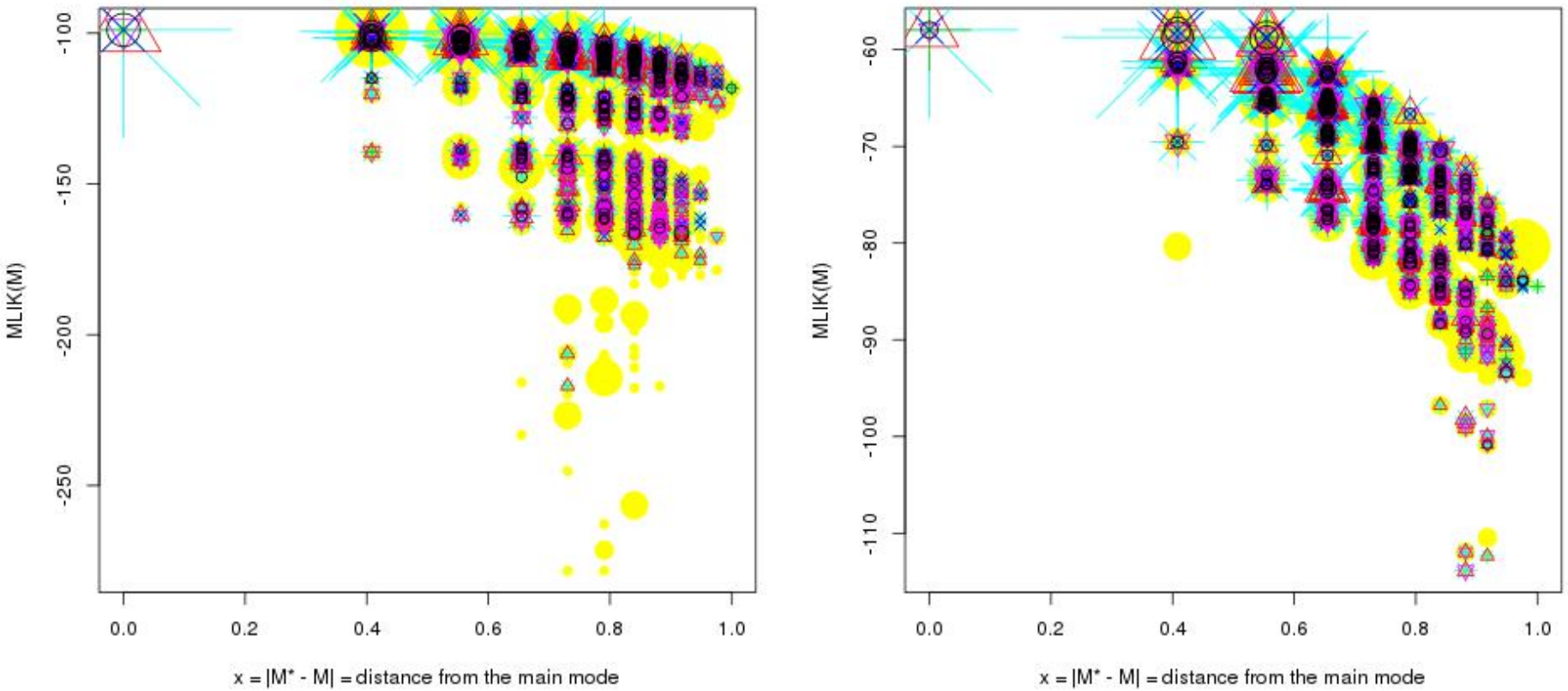



Figure: MLIK against distance from mcmc posterior mode

# Results

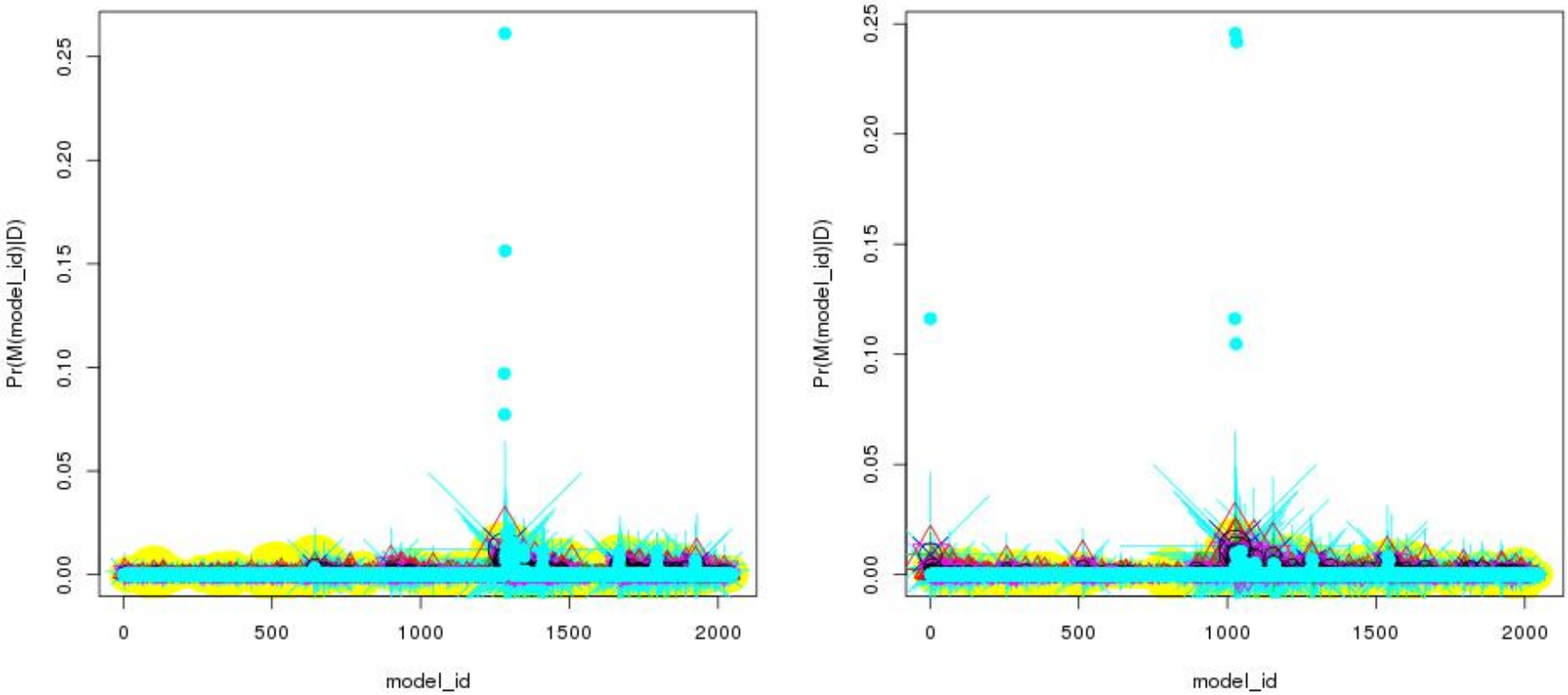



Figure: $\Pr(\mathbb{M}|\mathbb{D})$ against model index

# Results

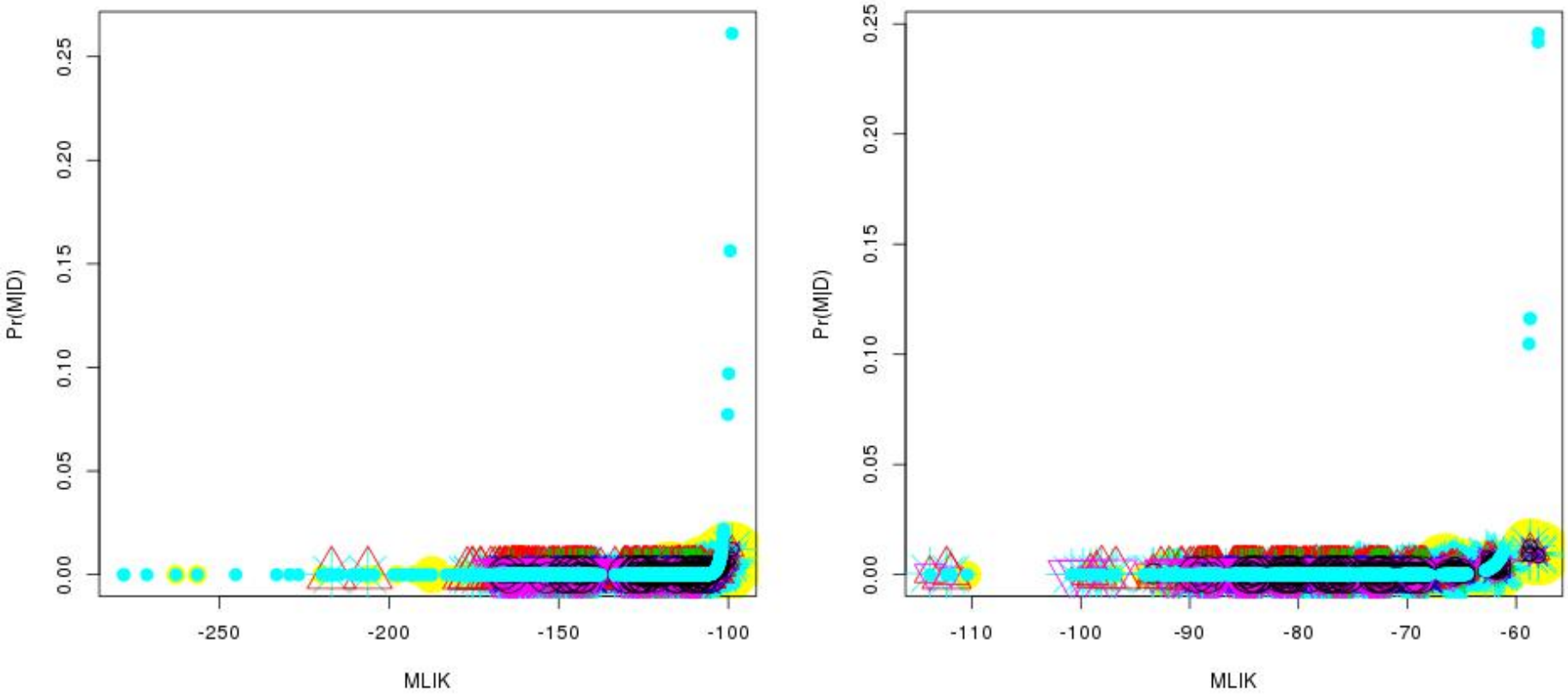



Figure: $\Pr(\mathbb{M}|\mathbb{D})$ against MLIK

# Results

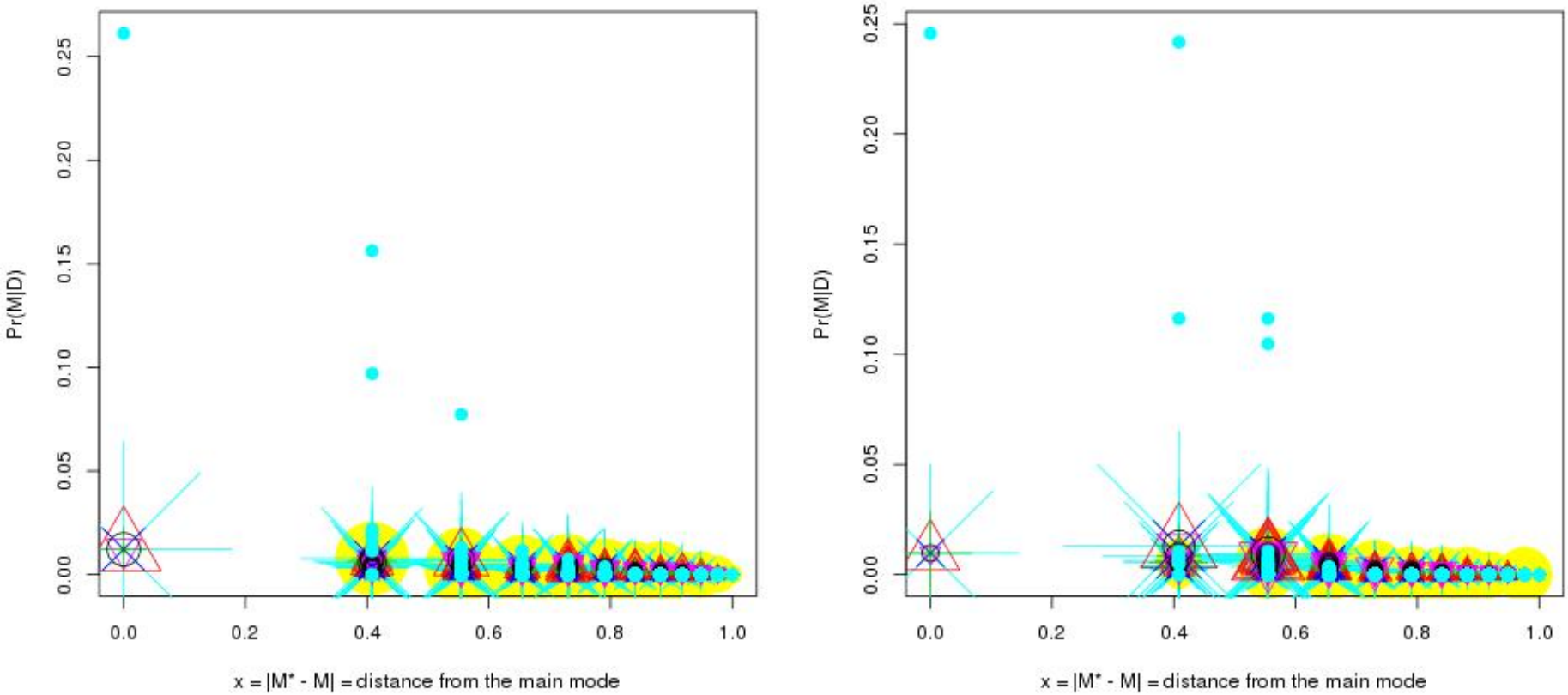



Figure: $\Pr(\mathbb{M}|\mathbb{D})$ against distance from MCMC posterior mode

# Results

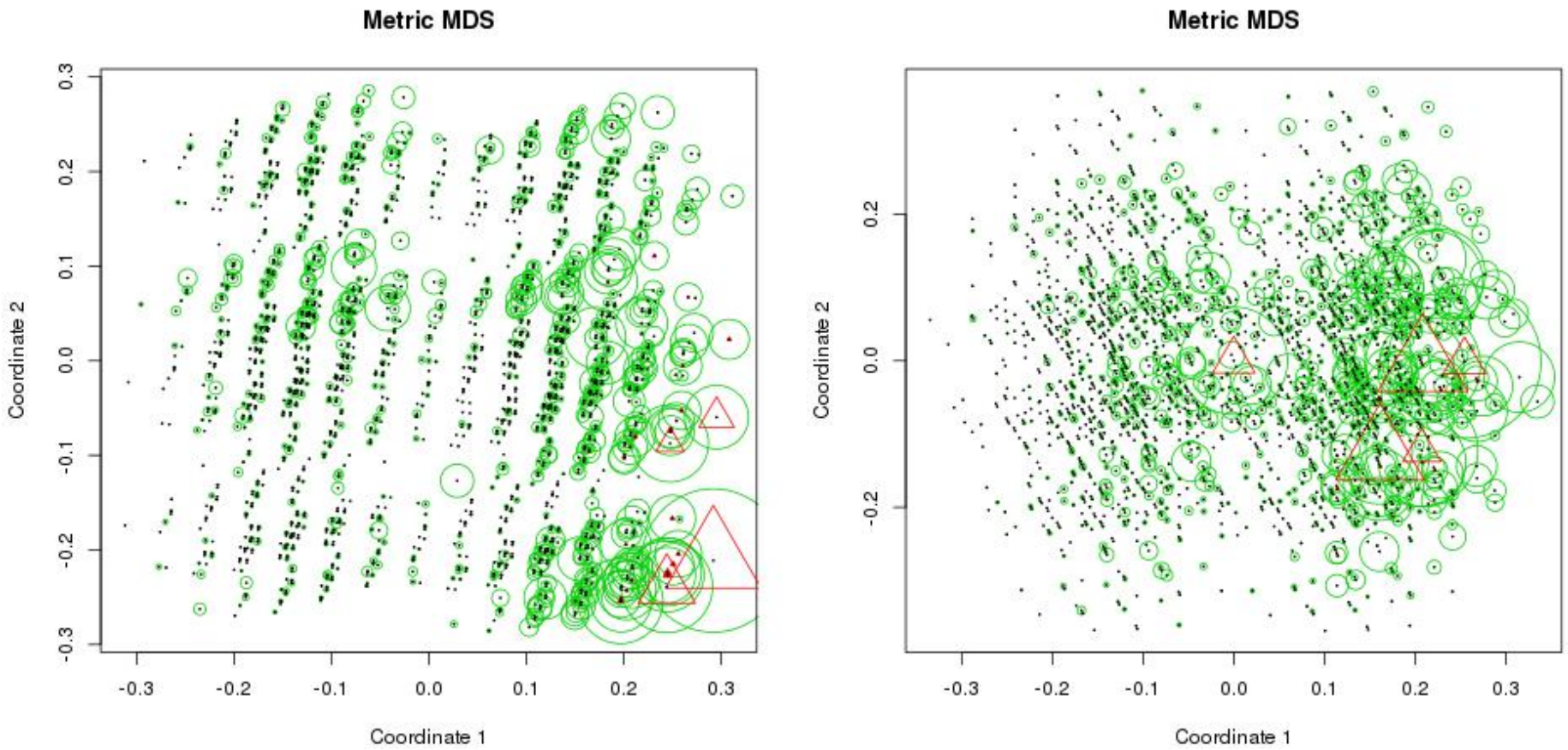



Figure: MDS plots with posterior modes

# Results

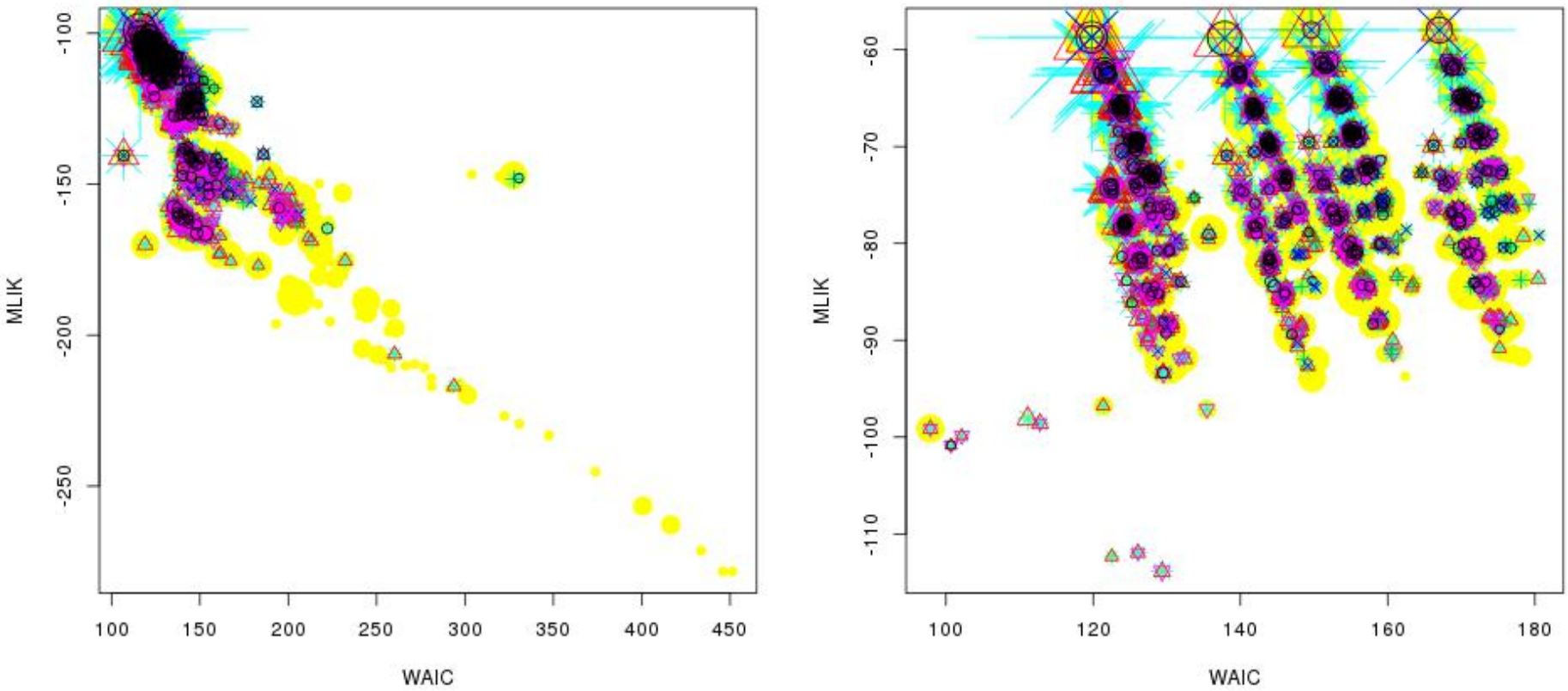



Figure: MLIK against WAIC

# Results

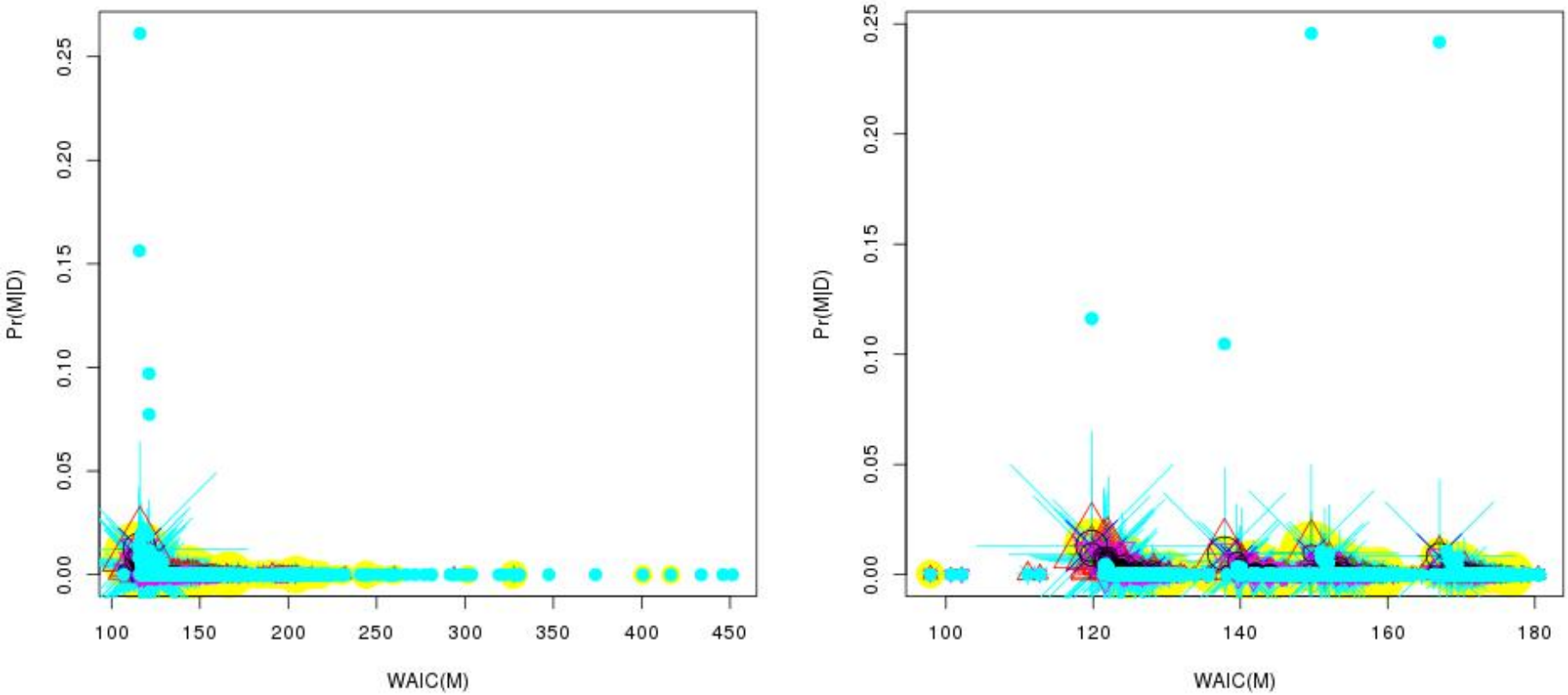



Figure: $\Pr(\mathbb{M}|\mathbb{D})$ against WAIC

# Some selected by WAIC models

| Ch | LkB | UkB | Solution | WAIC | MLIK | kTIME |
|---|---|---|---|---|---|---|
| 1 | 6485 | 6485 | 00101111110 | 97 | -99 | 19266691 |
| 1 | 5906 | 5907 | 10001000001 | 60 | -34 | 19049891 |
| 1 | 4851 | 4852 | 01001111110 | 78 | -81 | 18950460 |
| 1 | 2572 | 2572 | 10100000000 | 203 | -137 | 20660607 |
| 1 | 4395 | 4396 | 10000010001 | 106 | -140 | 21329179 |

**Table 1. Some results**

# Further discussion

- We see some heterogeneity of results, thus in order to achieve a stationary solution along the genome should we address one of the following:
  - Apply assessment of the subsets of the best solutions from different regions
  - Continue adding covariates which might be lacking
  - Somehow combine the strategies above
- The space of models $\Omega_{\mathbb{M}}$ can be extremely large for a large number of covariates to select, thus in order to achieve meaningful results should we
  - Carry out expert assessment of the found solutions in space $\Omega_{\mathbb{M}}$ a posteriori
  - Predefine biologically meaningful constraints $\mathbb{M} \in \Upsilon_{\mathbb{M}} \subset \Omega_{\mathbb{M}}$ before the search using standard mathematical modeling tools and thus limit down the search space
  - Combine the strategies above by both including constraints and carrying out a posteriori filtering of the results

# Concluding remarks

- We suggest using a model based approach for inference on methylation pattern along the genome
- We benefit of capturing local spatial correlation
- We suggest using different covariates to improve precision of inference
- We choose the combination of covariates optimally in order to reduce the amount of false positive and false negative discoveries
- Approach might be computationally expensive, thus efficient numerical algorithms are applied and/or developed
- Our model selection algorithm can be easily applied within any Bayesian variable selection, where marginal log likelihoods of the models are accessible including any of the observation and latent models within INLA package

# References

## REFERENCES


[1] C. Becker, J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt, and D. Weigel. Spontaneous epigenetic variation in the arabidopsis thaliana methylome. *Nature*, 480 (7376):245–249, 2011.

[2] D. Denilson, B. Mallick, and A. Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Sosciety*, 60(2):333–350, 1998.

[3] M. Denis and N. Molinary. Free knot splines with rjmcmc for logistic models and treshold selection. *JP Journal of Biostatistics*, 5(1):17–34, 2011.

[4] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014. ISSN 0960-3174. . URL http://dx.doi.org/10.1007/s11222-013-9416-2.

[5] E. I. George and R. E. Mcculloch. Approaches for bayesian variable selection. *Statistica Sinica*, pages 339–374, 1997.

# The End.

# Thanks!