Variable selection in logistic regression with a latent Gaussian field models for analysis of epigenetic data

Hubin A.A.,Storvik G.O.

Grini P.E., Lingjærde O.C., Butenko M.A.

University of Oslo

aliaksah@math.uio.no

UiO : Universitetet i Oslo

Third Semester Meeting

12.01.2016

Overview

1. Project plan
2. Introduction and biological motivation
3. The model
4. Inference
5. INLA
6. MCMC with mode jumping
7. Results
8. Conclusions

Courses and Teaching

Courses planned and completed

Term Planed	Term Done	Code	Result	Credits
Autumn 2014	Autumn 2014	STK9011	Pass	10.0
Autumn 2014	Autumn 2014	STK9021	Pass	10.0
Autumn 2014	Autumn 2014	STK9200	Pass	10.0
Spring 2015	Spring 2016	MNSES9100	-	5.0

Additionally NORINT 0110 is passed, NORA 0120 is taken in Spring 2016

Teaching obligations

Term	Code	Obligations
Spring 2015	STK2130	Plenaries, Tutorials, 1 assignment
Autumn 2015	STK3100/4100	Plenaries, 2 assignments
Autumn 2015	MAT1100	80 exam papers
Spring 2016	STK2130	Plenaries, Tutorials, 1 assignment

Presentations

Talks and posters

Date	Event	Topic
2015-02-09	SRI seminar	Statistics for Epigenetics
2015-05-29	Klækken Workshop, Klækken	On model selection in hidden Markov models with covariates
2015-10-30	Norbis Annual Meeting, Rosendal	Variable selection in binomial regression with a latent Gaussian field models for analysis of epigenetic data
2015-12-13	CMStatistics, London	Variable selection in binomial regression with a latent Gaussian field models for analysis of epigenetic data

Additionally two presentations at CELS meetings and 1 talk at the Statistics for Genomics discussion group were performed.

Articles

Articles

Planned	**Name**
Spring 2016	Efficient mode jumping MCMC for Bayesian model selection in GLM with a random effect models
Autumn 2016	Variable selection in logistic regression with a latent Gaussian field models for analysis of epigenetic data
Spring 2017	On model selection in hidden Markov models with covariates
Autumn 2017	To be decided

Introduction

- More precise estimation of the **methylation probability** of locations, which is represented by a number a binary events for all reads per given location
- Discovery of methylated and unmethylated regions and corresponding **local** and **global** structures:
  - Represented by nucleotides sequences patterns (**CPG-islands**)
  - Represented by such structures as **genes on the whole**, **promoters**, **coding regions** and their sequences
- Finding **covariates** (location within the gene, genetic structure, etc.) significantly influencing methylation patterns along the genome
- Linking genetic and epigenetic data to **phenotypic responses** (levels of **expression** of genes, presence of **transposons**, etc.) in a **statistically significant way**

Data visualization

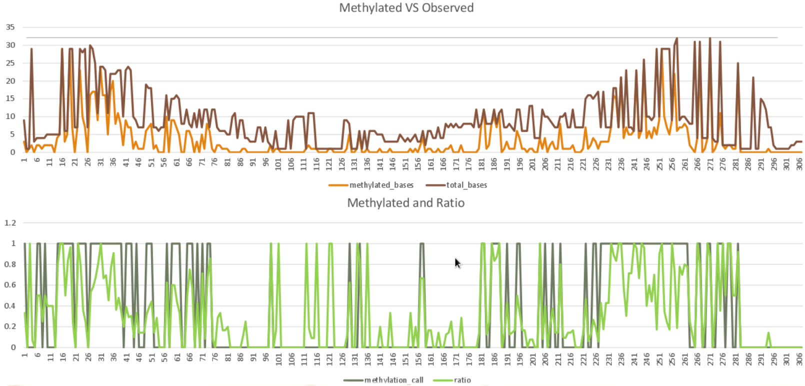


Figure: Total reads and methylated reads for some part of the genome

The model: Hierarchical Bayesian Model

The model: Logistic Regression With a Gaussian Latent Field Model (*Logistic Regression With a Random Effect Model*)

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

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

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

$$\epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{4}$$

- $y_t \in \{1, \dots, T\}$ is the number of methylated reads per loci t
- $n_t \in \mathbb{N}$ is the total number of reads per loci t
- $\beta_i \in \mathbb{R}, i \in \{0, \dots, M\}$ are regression coefficients of the covariates of the model
- δ_t is a Gaussian random effect of $AR(1)$ type with a parameter $\rho \in \mathbb{R}$
- ϵ_t is the error term of $AR(1)$

The model

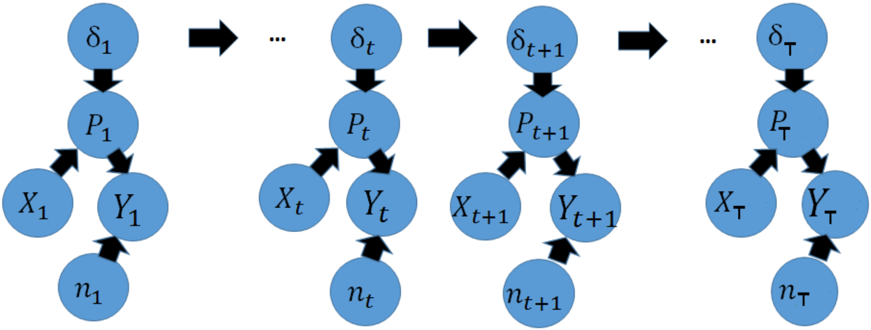


Figure: The model

T is extremely large $\Rightarrow$ Big Data

The model: Hyper-parameters of the model

We use a fully Bayesian approach, hence specify priors

$$\beta_i \sim N(\mu_\beta, \sigma_\beta^2) \tag{5}$$

$$\begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} \sim N_2(\mu_{\rho,\epsilon}, \Sigma_{\rho,\epsilon}) \tag{6}$$

- $\psi_1 = \log \frac{1}{\sigma_\epsilon^2}(1 - \rho^2)$ and $\psi_2 = \log \frac{1+\rho}{1-\rho}$ are scaled hyper-parameters of the latent model

The model: Model Selection

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

$$p_t = \frac{e^{\gamma_0\beta_0 + \sum_{i=1}^{N_\gamma} \gamma_i\beta_i X_{t,i} + \delta_t}}{1 + e^{\gamma_0\beta_0 + \sum_{i=1}^{N_\gamma} \gamma_i\beta_i X_{t,i} + \delta_t}} \tag{7}$$

$$\beta_i | \gamma_i \sim \mathbb{I}(\gamma_i = 1) N(\mu_\beta, \sigma_\beta^2) \tag{8}$$

$$\gamma_i \sim Binom(1, q) \tag{9}$$

- $\gamma_i \in \{0, 1\}, i \in \{0, ..., N_\gamma\}$ are latent indicators, defining if covariate i is included into the model
- q is the prior probability of including any covariate into the model, which corresponds to the spike and slab model

Inference on the model

Let:

- $\mathbb{M} = \vec{\gamma}$ be further addressed as simply a model
- $\Theta|\mathbb{M}$ define parameters conditioned on fixed models
- $\exists 2^{N_\gamma + 1}$ different models

Goals:

- $\Pr(\mathbb{M}, \Theta|\mathbb{D})$ posterior distribution of parameters and models
- $\Pr(\mathbb{M}|\mathbb{D})$ 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** $\Pr(\mathbb{M},\Theta|\mathbb{D}) = \Pr(\Theta|\mathbb{M},\mathbb{D})\Pr(\mathbb{M}|\mathbb{D})$
- $\Pr(\Theta|\mathbb{M},\mathbb{D})$ 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 $\Omega_{\mathbb{M}}$

INLA overview

Assume
Observation model: $\pi(y|\eta)$
Parameter model: $\pi(\eta|\upsilon) \sim N_u(\mu(\upsilon), Q(\upsilon))$
Hyperparameter: $\upsilon \sim f(\upsilon)$

The models are assumed to satisfy some properties:

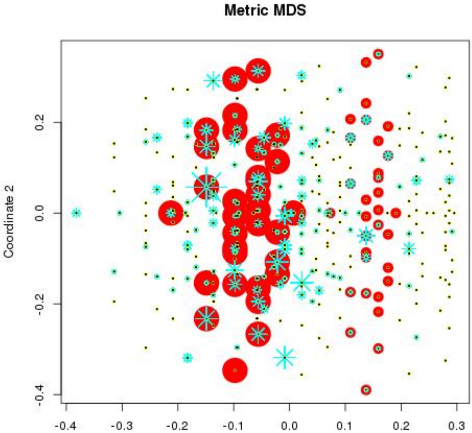
- The parameter can can be of big size but with a sparse precision matrix
- The dimension of the hyperparameter vector υ is relatively small
- Laplace approximation method of the posterior density can be used

INLA efficiently calculates:

- The marginal posterior distribution of parameters which can be summarized by means, variances and quantiles
- Model selection criteria DIC, WAIC, MLIK (exactly $\log \Pr(\mathbb{D}|\mathbb{M} = M)$)
- Predictive measures (CPO, PIT)

Model selection problems

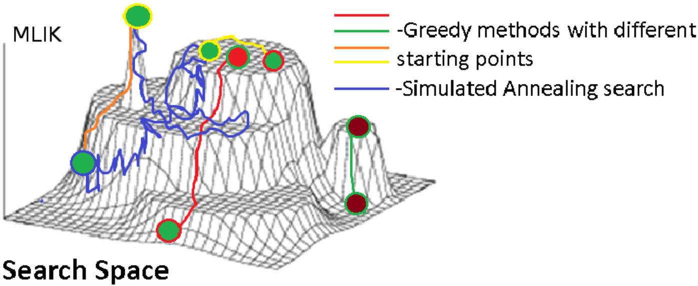
- Proceed with **efficient exploration of** $\mathbb{V}$ in the subspace of $\Omega_{\mathbb{M}}$ **to estimate** $\Pr(\mathbb{M} = M|\mathbb{D})$, $\underset{M\in\Omega_{\mathbb{M}}}{\operatorname{argmax}} \Pr(\mathbb{M} = M|\mathbb{D})$, **and** $\underset{M\in\Omega_{\mathbb{M}}}{\operatorname{argmax}} \text{WAIC}(M)$
- Main challenges are **multimodality** in $\Omega_{\mathbb{M}}$ and its **size**



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

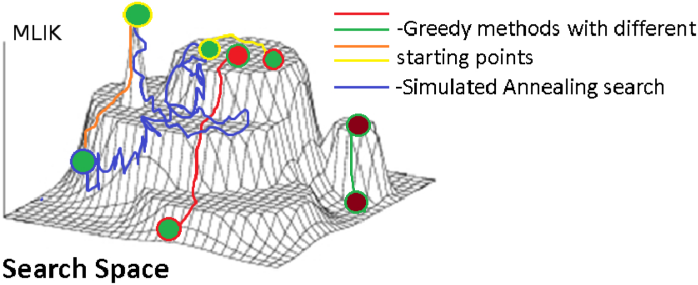
Main challenges are multimodality in $\Omega_{\mathbb{M}}$ and its size.

- 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


MLIK
-Greedy methods with different starting points
-Simulated Annealing search
Search Space

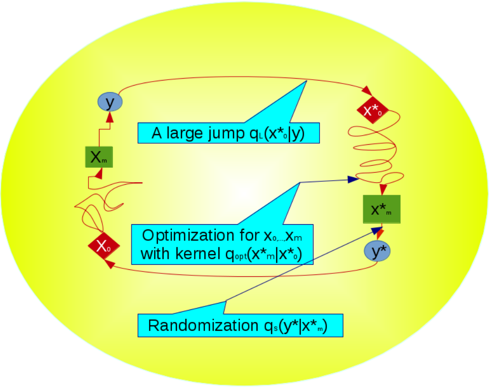
Treatments for multimodality to explore $\mathbb{V} \subset \Omega_{\mathbb{M}}$

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


MLIK
-Greedy methods with different starting points
-Simulated Annealing search
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.



y
x^*_0
A large jump $q_L(x^*_0|y)$
X_m
x^*_m
Optimization for $x_0,\dots,x_m$ with kernel $q_{opt}(x^*_m|x^*_0)$
X_0
y^*
Randomization $q_s(y^*|x^*_m)$

Figure: Locally optimized with randomization proposals

MCMC balance with mode jumping proposals

forward move	**comment**	**backward move**
$y \sim \pi(y)$	initial state	$y^* \sim \pi(y^*)$
$x_0^* \sim \mathrm{q_L}(x_0^*\|y)$	large jump	$x_0 \sim \mathrm{q_L}(x_0\|y^*)$
$x_m^* \sim \mathrm{q_{opt}}(x_m^*\|x_0^*)$	optimization	$x_m \sim \mathrm{q_{opt}}(x_m\|x_0)$
$y^* \sim \mathrm{q_s}(y^*\|x_m^*)$	randomization	$y \sim \mathrm{q_s}(y\|x_m)$
$(x^*, y^*) \sim \mathrm{w}(x^*, y^*\|y)$	thus	$(x, y) \sim \mathrm{w}(x, y\|y^*)$
$x\|y, x^*, y^* \sim \mathrm{h}(x\|y, x^*, y^*)$	choose	$x^*\|y, x, y^* \sim \mathrm{h}(x^*\|y, x, y^*)$

$$\blacktriangleright\ \pi(y,x)A(y,x;y^*,x^*) = \pi(y)\mathrm{w}(y^*,x^*|y)\mathrm{h}(x|y^*,x^*,y)r_m(y,x;y^*,x^*)$$

$$= \pi(y)\mathrm{w}(y^*,x^*|y)\min\left\{1, \frac{\pi(y^*)\mathrm{w}(y,x|y^*)\mathrm{h}(x^*|y,x,y^*)}{\pi(y)\mathrm{w}(y^*,x^*|y)\mathrm{h}(x|y^*,x^*,y)}\right\}$$

$$= \pi(y^*)\mathrm{w}(y,x|y^*)\mathrm{h}(x^*|y,x,y^*)\min\left\{\frac{\pi(y)\mathrm{w}(y^*,x^*|y,x)\mathrm{h}(x|y^*,x^*,y)}{\pi(y^*)\mathrm{w}(y,x|y^*)\mathrm{h}(x^*|y,x,y^*)}, 1\right\}$$

$$= \pi(y^*,x^*)A(y^*,x^*;y,x)\blacktriangleleft \qquad (10)$$

Application of MCMC with mode jumping proposals

Let $y = \mathbb{M}_j$, $y^* = \mathbb{M}_k$, $x^* = \{\mathbb{M}_{k_0}\},...\{\mathbb{M}_{k_{K-1}}\}$, and $x = \{\mathbb{M}_{j_0}\},...\{\mathbb{M}_{j_{K-1}}\}$ and h(|) be in the form (12) then (10) becomes:

$$r_m(\mathbb{M}_j,\mathbb{M}_k) = \min\left\{1, \frac{\Pr(D|\mathbb{M}_k)\Pr(\mathbb{M}_k)\mathrm{q_s}(\mathbb{M}_j|\mathbb{M}_{j_{K-1}})\mathrm{q_s}(\mathbb{M}_{k_{K-1}}|\mathbb{M}_k)}{\Pr(D|\mathbb{M}_j)\Pr(\mathbb{M}_j)\mathrm{q_s}(\mathbb{M}_k|\mathbb{M}_{k_{K-1}})\mathrm{q_s}(\mathbb{M}_{j_{K-1}}|\mathbb{M}_j)}\right\}. \tag{11}$$

with

$$\begin{aligned}\mathrm{h}(\mathbb{M}_{j_0},...,\mathbb{M}_{j_{K-1}}|\mathbb{M}_k,\mathbb{M}_j,\mathbb{M}_{k_0},...,\mathbb{M}_{k_{K-1}}) &= \mathrm{q_L}(\mathbb{M}_{j_0}|\mathbb{M}_k)\times \\ &\times\prod_{i\in\{1,...,K-2\}}\mathrm{Q}\left(\mathbb{M}_{j_i}|\mathbb{M}_{j_{i-1}}\right)\mathrm{q_s}\left(\mathbb{M}_{j_{K-1}}|\mathbb{M}_j\right)\end{aligned} \tag{12}$$

where Q(.|.) is the transition kernel of the local optimization algorithm and $\mathrm{q_s}(.|.)$ is the kernel of randomization at the end.

MCMC with mode jumping proposals

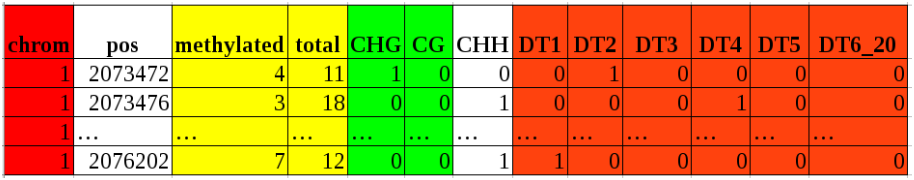
Notice that

Locally annealed, **locally optimized**, **locally simulated** and **locally multiple try simulated** proposals and **their combination** are all **of this type of extension of the original space** and therefore their detailed balanced equation is proven in (10).

Also notice

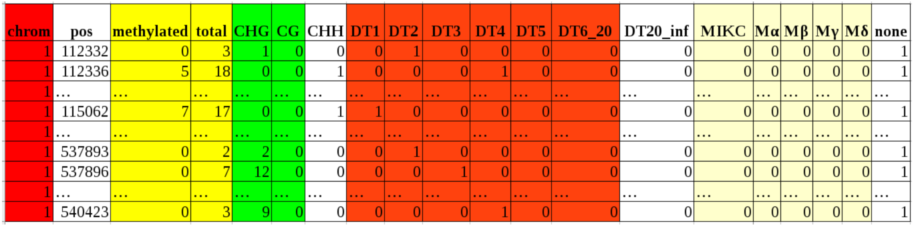
Also note that within this setting of locally optimized MCMC we get an **alternative MCMC based approximations** for posterior probabilities of the **models**, namely $\tilde{\Pr}(\mathbb{M} = M|\mathbb{D}) = \frac{\sum_{i=1}^{W} \mathbb{I}(M_i = M)}{W} \xrightarrow[W\to\infty]{d} \Pr(\mathbb{M} = M|\mathbb{D})$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

Variables. Data



chrom	pos	methylated	total	CHG	CG	CHH	DT1	DT2	DT3	DT4	DT5	DT6_20
1	2073472	4	11	1	0	0	0	1	0	0	0	0
1	2073476	3	18	0	0	1	0	0	0	1	0	0
1	…	…	…	…	…	…	…	…	…	…	…	…
1	2076202	7	12	0	0	1	1	0	0	0	0	0

Figure: Variables and data addressed. Small example (9 variables)



chrom	pos	methylated	total	CHG	CG	CHH	DT1	DT2	DT3	DT4	DT5	DT6_20	DT20_inf	MIKC	Mα	Mβ	Mγ	Mδ	none
1	112332	0	3	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1	112336	5	18	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
1	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1	115062	7	17	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1	537893	0	2	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
1	537896	0	7	12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
1	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1	540423	0	3	9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1

Figure: Variables and data addressed. Large example (14 variables)

9 variables. Comparisons of strategies on 40000 iterations

C	LkB	UkB	optim	CV	DN	ID	WAIC	MLIK
1	20734	20749	no-int	0.56	0.3775	320	-127.6	179.1
1	20734	20749	no-div	0.65	0.3832	320	-127.6	179.1
1	20734	20749	greed	0.77	0.4507	320	-127.6	179.1
1	20734	20749	mcmc	0.90	0.4508	320	-127.6	179.1
1	20734	20749	SA-1	0.91	0.4508	320	-127.6	179.1
1	20734	20749	SA-2	0.88	0.4507	320	-127.6	179.1
1	20734	20749	mix	0.99	0.4508	320	-127.6	179.1

Table 1. Comparison of strategies

$$\mathrm{DN} = \exp\left(K\right)\sum_{M' \in \mathbb{V}} exp((\log \Pr(\mathbb{D}|\mathbb{M} = M') + \log \Pr(\mathbb{M} = M'))$$

$$\mathrm{ID} = \mathrm{toDEC}\left(\underset{M \in \mathbb{V}}{\operatorname{argmax}} \Pr(\mathbb{M} = M|D)\right) \Leftrightarrow \underset{M \in \mathbb{V}}{\operatorname{argmax}} \Pr(\mathbb{M} = M|D) = \mathrm{toBIN}(\mathrm{ID})$$

$\mathrm{CV} = \frac{||\mathbb{V}||}{||\Omega_{\mathbb{M}}||}$, $||.||$ - cardinality of a set

Results. Legend

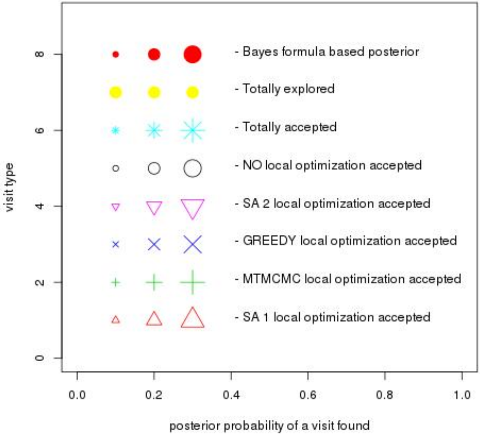
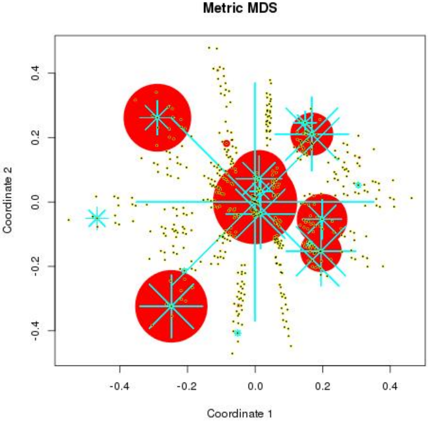


Figure: Notation used

Results. 9 variables

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



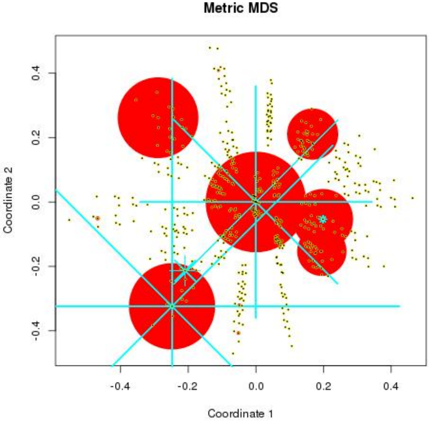


Figure: MDS plots with posterior modes of all found solutions

Results. 9 variables

Mode jumping proposals - better MCMC approximations. Modes have overestimated probabilities (right figure) when some are missed

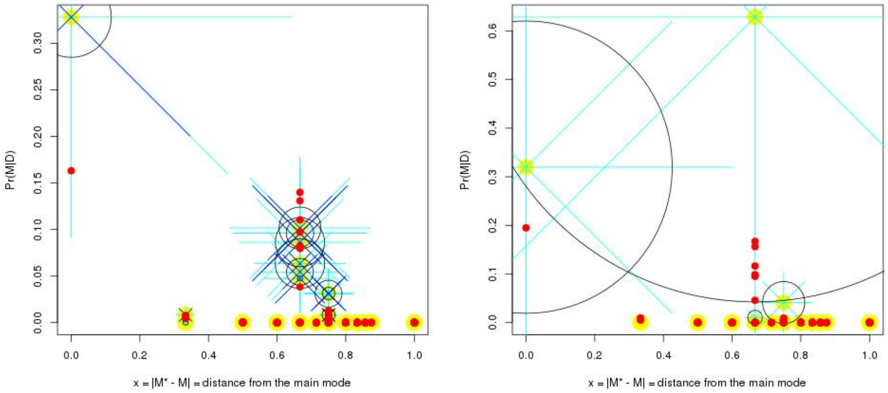
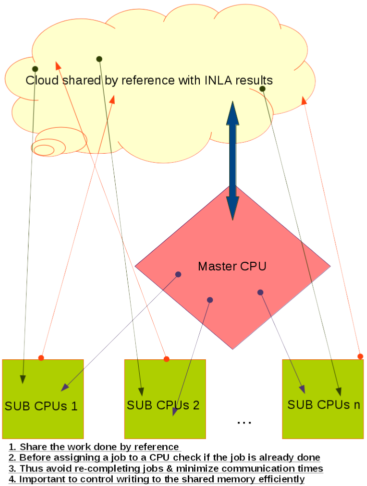


Figure: Posterior probability versus distance from the global mode

Multicore and shared memory issues



1. Share the work done by reference
2. Before assigning a job to a CPU check if the job is already done
3. Thus avoid re-completing jobs & minimize communication times
4. Important to control writing to the shared memory efficiently

Figure: Multiprocessing architecture

Results. Combination of optimizers

We now apply a mixture of local optimizers with greedily optimized frequences or kernel of their appearance learned during the burn-in

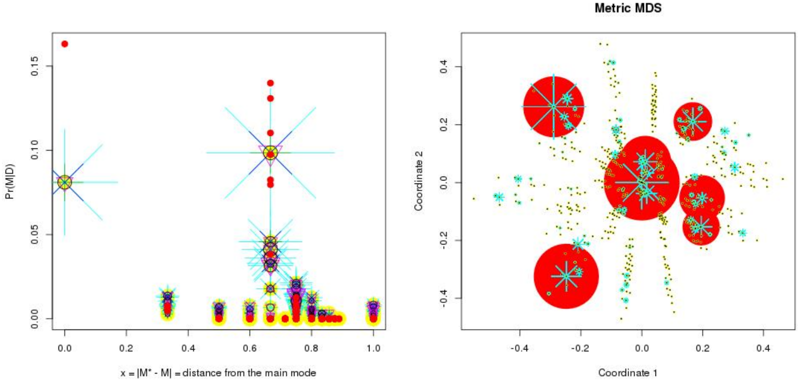


Figure: Combination of locally optimized proposals

Results. Now 14 decision variables

Apply a mixture in the exponentially larger space of variables

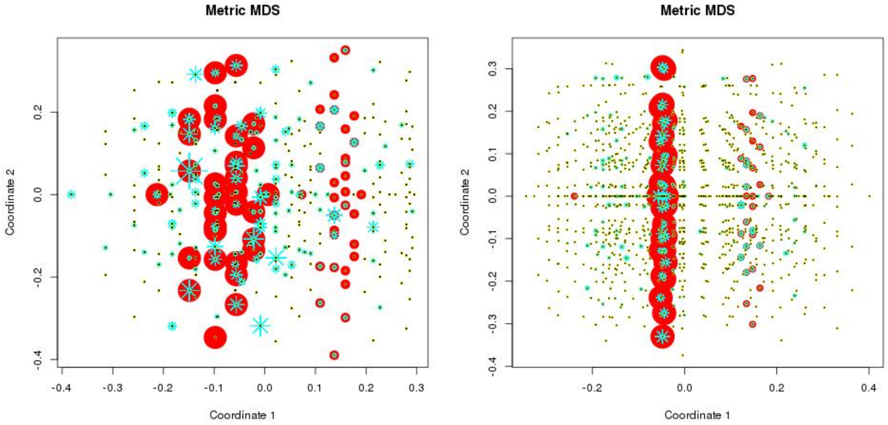


Figure: MDS plots with posterior modes of best 1024 solutions

Results. Now 14 decision variables

WAIC is yet another story...

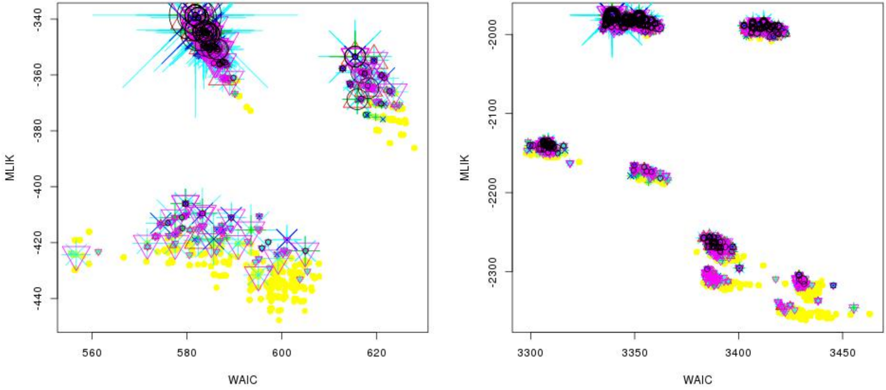


Figure: MLIK against WAIC

Selected models. Now 14 decision variables, 10^6 iterations

Missing a few modes dramatically influences the results

C	LkB	UkB	optim	CV	DN	ID	WAIC	MLIK
1	112	115	no-int	0.49	0.0379	14367	-127.6	179.1
1	112	115	mix	0.80	0.3832	14367	-127.6	179.1

Table 2. First data set. Comparison of strategies

C	LkB	UkB	Solution	WAIC	MLIK
1	112	115	11100000011111	581.8	-338.6
1	112	115	10000111100011	555.6	-423.8
1	537	540	11100000011111	3339.1	-1975.7
1	537	540	10101101100001	3297.8	-2150.8

Table 3. Some results. Two data sets. Mixture of improvements

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 variables to improve precision of inference
- We carry out efficient choice of the subsets of these variables with respect to posterior marginal model probability and other criteria by means of mode jumping MCMC strategies
- Approach might be computationally expensive, since the nature of such a search is NP-hard, thence we efficiently address both mode jumping and parallel computation providing reasonably fast communication of the central processing units involved
- Model selection procedure developed is not problem specific and can be easily adopted to any problem where marginal likelihoods of the models are available. In particular it gives a general model selection tool within a popular INLA approach

References

References.

[1] R. Bonneville and V. X. Jin. A hidden markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor α target genes. *Bioinformatics*, 29(1): 22–28, 2013.

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

[3] W. H. W. Jun S. Liu, Faming Liang. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

[4] 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.

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

[6] H. Tjelmeland and B. K. Hegstad. Mode jumping proposals in mcmc. *SCANDINAVIAN JOURNAL OF STATISTICS*, 28:205–223, 1999.

The End.

Thanks!