MCMC methods for gene expression profiling via Bayesian variable selection

Manuela Zucknick[12], Sylvia Richardson[2]

[1]Dept of Biostatistics, German Cancer Research Centre, Heidelberg

[2]Centre for Biostatistics, Imperial College London

Specificity of the context of application

- Many more covariates (thousands of genes) than samples (~ 100): large p, small n paradigm.
- Gene prediction: building molecular profiles based on gene expression which can characterise different phenotypes (e.g. clinical outcomes).
- Many such questions can be framed in a regression set-up, with the focus on variable selection.
- Here: binary classification $\longrightarrow$ probit/logistic regression
- Sparseness: Out of the thousands of genes usually only a few are expected to be related to the response.
- Complex dependence structure between genes linked to underlying biological pathways and networks.

Difficulties

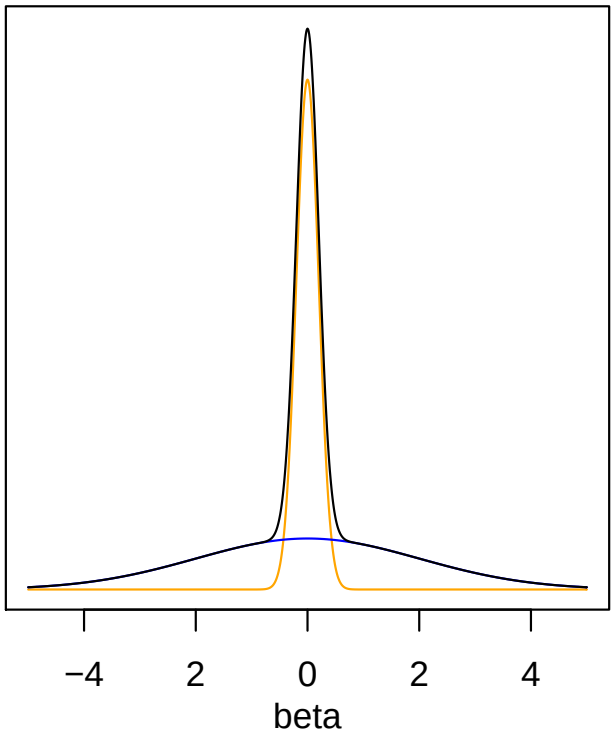
- In the "large p, small n" context, regression is an ill-conditioned problem, with a multi-modal posterior distribution over the model space, i.e. many alternative models having similar explanatory power
- The model space becomes huge, of size 2^p (when no interactions included) and full exploration is unfeasible
- $\longrightarrow$ Use of MCMC methods as stochastic search algorithms

Motivation for Bayesian variable selection

BVS model with indicator variable $\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$

Shape of prior to encourage sparsity:

- spike in zero (variable exclusion),
- heavy tails (variable inclusion).



Examples:

- Normal mixture prior

  (George and McCulloch 1997):

  $$\beta_i|\gamma_i \sim (1-\gamma_i)N(0, \sigma^2 v_{0\gamma_i}) + \gamma_i N(0, \sigma^2 v_{1\gamma_i})$$

- Conditional model

  (Holmes and Held 2006):

  β_i only defined and estimated if $\gamma_i = 1$

Motivation for Bayesian variable selection

- A prior on the model space can be specified via a prior $p(\gamma)$.
  Here: Binomal prior $p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i} (1-\pi_i)^{1-\gamma_i}$
- This approach was taken by George and McCulloch (1993, 1997) and many subsequent authors (Clyde, 1999; Brown et al 1998, 2002, Kohn et al 2001, Jasra et al 2005, Cui and George 2006, ...).
- Much of the work on BVS has been done in the linear/probit regression context, where typically, the choice of a conjugate prior for the regression vector β is used, allowing β to be integrated out.
- In other cases (logistic regression), it is important to update the covariate indicator and the regression coefficients jointly as these are typically strongly correlated, especially when the covariates X are non-orthogonal.
- In the binary case, auxiliary variable representation is called upon: either for the probit (Albert and Chib 1993, Lee et al 2003) or the logit link (Holmes and Held 2006)

Logistic BVS model

Probit regression

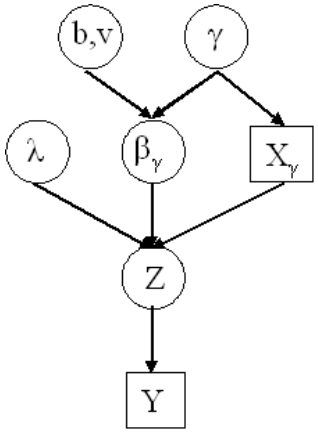
(Albert and Chib 1993)

$$
\begin{aligned}
y_{\gamma j} &= \begin{cases} 1 & \text{if } z_{\gamma j} > 0 \\ 0 & \text{otherwise} \end{cases} \\
z_{\gamma j} &= x_{\gamma j}\beta_\gamma + \epsilon_j \\
\epsilon_j &\sim N(0,1) \\
\\
\beta_\gamma &\sim N(b_\gamma, v_\gamma) \\
\gamma &\sim p(\gamma)
\end{aligned}
$$

Logistic regression

(Holmes and Held 2006)

$$
\begin{aligned}
y_{\gamma j} &= \begin{cases} 1 & \text{if } z_{\gamma j} > 0 \\ 0 & \text{otherwise} \end{cases} \\
z_{\gamma j} &= x_{\gamma j}\beta_\gamma + \epsilon_j \\
\epsilon_j &\sim N(0,\lambda_j) \\
\lambda_j &= (2\phi_j)^2 \\
\phi_j &\sim \text{KS, independently} \\
\beta_\gamma &\sim N(b_\gamma, v_\gamma) \\
\gamma &\sim p(\gamma)
\end{aligned}
$$



γ subscripts indicate that variable is only defined for those i with $\gamma_i = 1$.

Here, we focus on the logistic regression model because of better interpretability in terms of odds ratios.

Benefits of auxiliary variable representation

- The ϵ_j have a **scale mixture of normals** form with marginal logistic distribution (Andrews and Mallows 1974)
- If the prior for β_γ is normal, $p(\beta_\gamma) = N(b_\gamma, v_\gamma)$, so is the posterior:

$$\begin{aligned} \beta_\gamma | z_\gamma, \lambda &\sim N(B_\gamma, V_\gamma) \\ B_\gamma &= V_\gamma(v_\gamma^{-1} b_\gamma + X_\gamma^T \lambda^{-1} z_\gamma) \\ V_\gamma &= (v_\gamma^{-1} + X_\gamma^T \lambda^{-1} X_\gamma)^{-1} \\ \lambda^{-1} &= \text{diag}(\lambda_1^{-1}, ..., \lambda_n^{-1}). \end{aligned}$$

Note that V_γ depends on λ, so needs to be recomputed at every sweep.

Typically, the hyper-parameters for the prior for β_γ are $b_\gamma = 0$ and either

- independence prior covariance $v_\gamma = c^2 I_{p_\gamma}$ **(our choice with $c^2 = 5$)**
- or g-prior covariance $v_\gamma = c^2 (X_\gamma^T X_\gamma)^{-1}$ (Bottolo and Richardson 2007)

MCMC sampler for logistic BVS

- Update $\{z, \lambda\}$ jointly given $\{\beta, \gamma\}$:

$$p(z, \lambda|\beta_\gamma, X_\gamma, y) = p(\lambda|z, \beta_\gamma, X_\gamma)p(z|\beta_\gamma, X_\gamma, y)$$

  (the second distribution on the RHS is a truncated logistic)

- Update $\{\beta, \gamma\}$ jointly given $\{z, \lambda\}$ with joint proposal

$$p(\gamma^*, \beta^*) = p(\beta^*|\gamma^*, z, \lambda, X)q(\gamma^*) = N(B_{\gamma^*}, V_{\gamma^*})q(\gamma^*)$$

  This is done via a Metropolis-Hastings step to update the current covariate set (defined by γ), with a subsequent update to β. The acceptance probability of the joint move is

$$\alpha = \min\left\{1, \frac{|V_{\gamma^*}|^{1/2}|v_\gamma|^{1/2}}{|V_\gamma|^{1/2}|v_{\gamma^*}|^{1/2}} \frac{\exp(0.5B'_{\gamma^*}V_{\gamma^*}^{-1}B_{\gamma^*})}{\exp(0.5B'_\gamma V_\gamma^{-1}B_\gamma)} \frac{\mathbf{p}(\gamma^*)\mathbf{q}(\gamma|\gamma^*)}{\mathbf{p}(\gamma)\mathbf{q}(\gamma^*|\gamma)}\right\}$$

- Note that α does not involve β_γ or β^*_γ, but only its posterior mean B_γ and variance $V_\gamma \rightarrow$ efficient updates

MCMC schemes for updating the covariate set indicator γ

- Vanilla sampler (add/delete move) (Holmes and Held, 2006)

  Select indicator variable γ_k at random and change state ($0 \rightarrow 1$ or $1 \rightarrow 0$):

$$\frac{p(\gamma^*)q(\gamma|\gamma^*)}{p(\gamma)q(\gamma^*|\gamma)} = \frac{q(\gamma|\gamma^*)\prod_{i=1}^{p}\pi_i^{\gamma_i^*}(1-\pi_i)^{1-\gamma_i^*}}{q(\gamma^*|\gamma)\prod_{i=1}^{p}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}} = \begin{cases} \frac{\pi_k}{1-\pi_k} & \text{if } \gamma_k = 0 \\ \frac{1-\pi_k}{\pi_k} & \text{if } \gamma_k = 1 \end{cases}.$$

  For large p, and dependent X covariates proposing to update one indicator variable at a time leads to **very slow mixing** (see comparison later).

- Other extreme: Propose to update the entire indicator vector, for example using a Gibbs sampling proposal, will be computationally very expensive (Lee 2003)

- Block update using blocks of (partially) correlated variables

  Idea: If we can **assume sparse dependence structure** in X, only those genes which are (partially) correlated need to be updated together.

  $\rightarrow$Define neighbours for each gene based on correlation or partial correlation

Block (or neighbourhood) proposal scheme

1. Select a gene i randomly.
2. Propose to update neighbouring genes $k \in N(i)$ together with selected gene.
3. Propose new values $\gamma_k^*, k \in \{i, N(i)\}$, for these genes by drawing from their univariate conditional distributions:

$$\begin{aligned} p(\gamma_k^*|\gamma_{-k}, z, X, \lambda) &\propto p(z|\gamma, X, \lambda)p(\gamma) \qquad (\text{with } \gamma = (\gamma_k^*, \gamma_{-k})) \\ &= N(0, (\lambda^{-1} - \lambda^{-1}X_\gamma V_\gamma X_\gamma' \lambda^{-1})^{-1}) \prod_{i=1}^{p} \pi^{\gamma_i}(1-\pi)^{1-\gamma_i} \end{aligned}$$

Normalising constant unknown $\longrightarrow$ compute for both $\gamma_k = 1$ and $\gamma_k = 0$.

How to determine blocks, i.e. estimate neighbourhood structure

1. Correlation ($Corr$) or partial correlation ($Pcor$) matrix: Estimate either using a shrinkage estimator (R library `corpcor`, Schäfer and Strimmer 2005).
2. Threshold C: Minimum size of absolute pairwise (partial) correlations for two variables to be declared neighbours. Here, the threshold is set to the Cth percentile of all absolute (partial) correlations ($Corr < C >$, $Pcor < C >$).

Need to balance the improvement in mixing with the extra computational burden of sampling a large number of indicators

Comparisons of block samplers with:

- AD - add/delete sampler
- $Full$ - full Gibbs sampler updating every variable by sampling from its full conditional distribution

Simulation examples A and B

Example A (25 replications)

- Similar to example 4.2 in George and McCulloch (1993)
- $p = 100 \times 5$ variables and $n = 100$ samples
- Variables are correlated by blocks of size 100
- $p^* = 5$ variables are related to response y (true model) via logistic link: $\beta = (2, 2, 2, 2, 2, 0, ..., 0)$

Example B (25 replications)

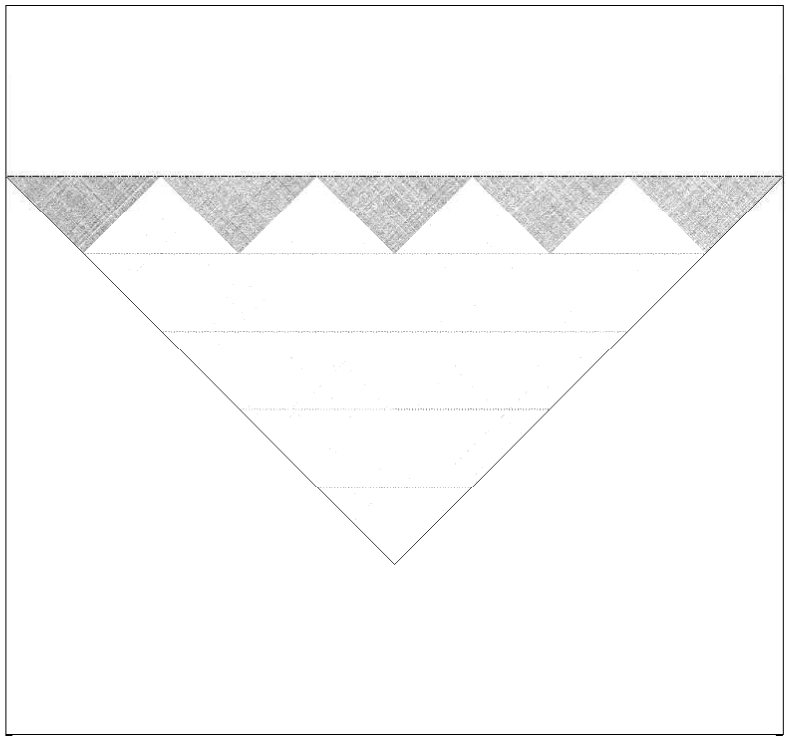
- Sample p genes from real gene expression data (Schwartz *et al.* 2002)
- $p = 500$ variables and $n = 104$ samples
- $p^* = 5$ variables are related to response y (true model) via logistic link: $\beta = (2, 2, 2, 2, 2, 0, ..., 0)$

– For both examples, prior probability for $\gamma_i = 1 : \pi = \frac{p^*}{p} = 0.01$

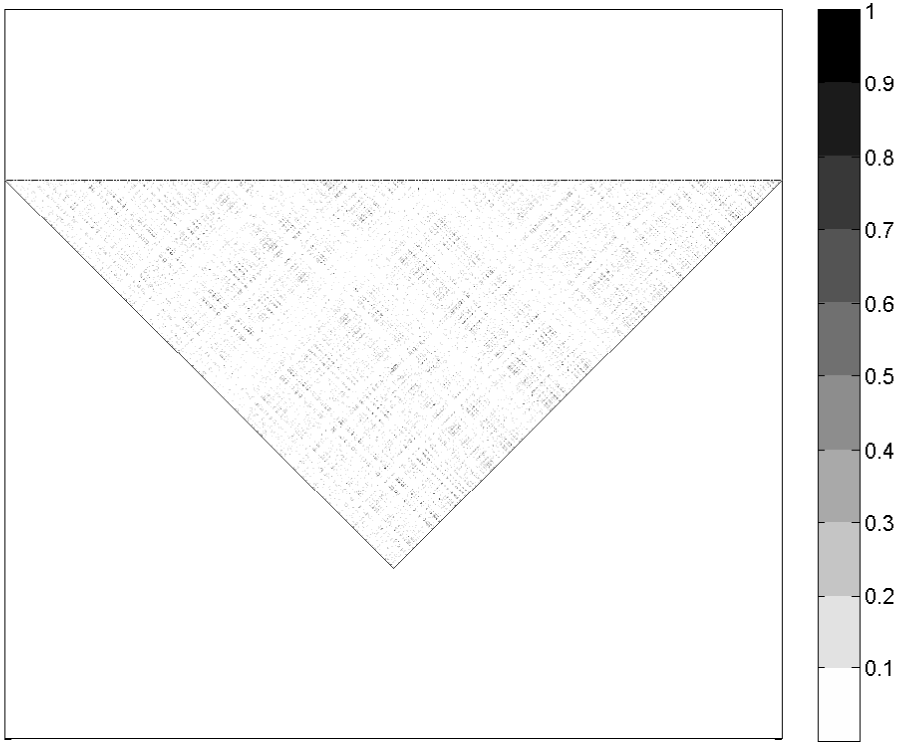
– Computing specs: Matlab, dual-core PC: CPUs @ 2.40GHz, 3.5GB RAM

Simulation examples A and B

Correlation structure - Example A



Correlation structure - Example B



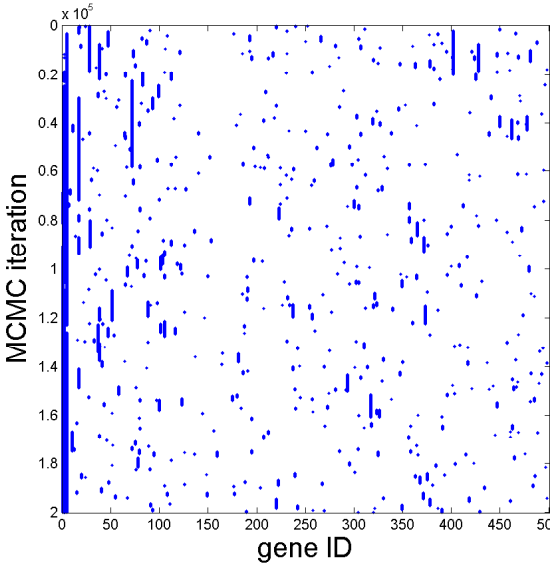
- Add/delete and block samplers:
  $T = 200,000$, $B = 50,000$
- Full Gibbs:
  $T = 90,000$, $B = 10,000$

- Add/delete and block samplers:
  $T = 250,000$, $B = 50,000$
- Full Gibbs:
  $T = 110,000$, $B = 10,000$

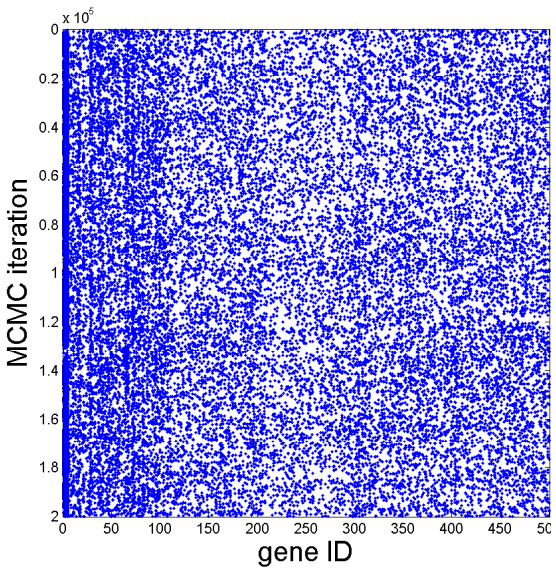
Simulation example A

Trace plots of γ vector show improved mixing for block update (results for one replicate)

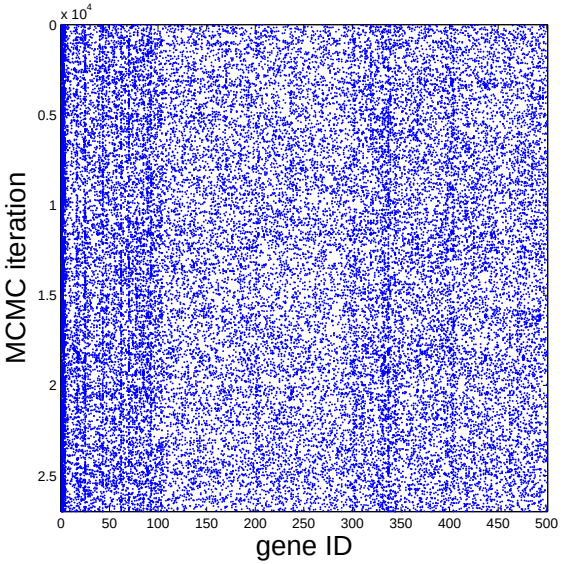
Add/delete sampler



Block (Pcor, $C = 90\%$)

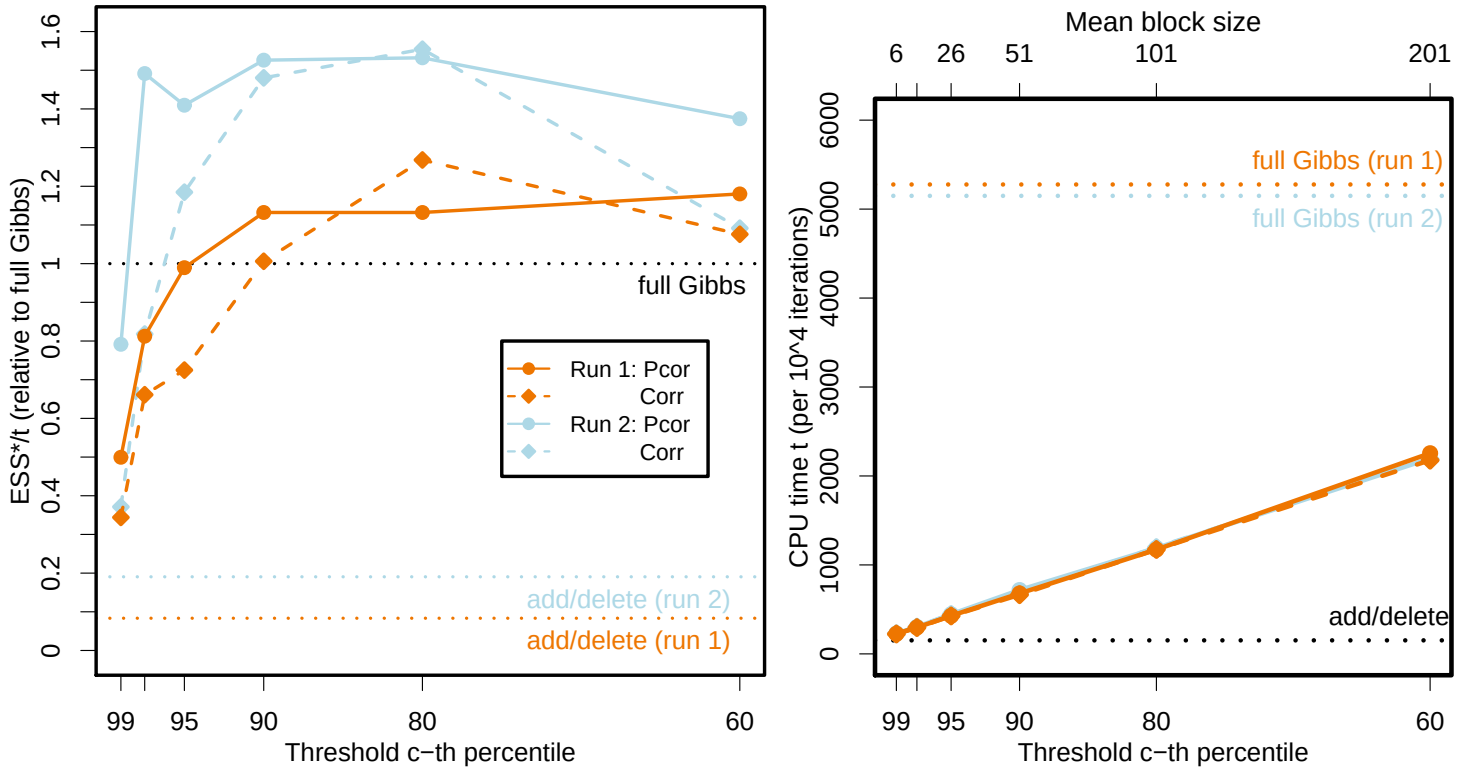


Full Gibbs



Simulation results

Relative median effective sample sizes (ESS^*/t) and CPU times t for various choices of block size threshold C (results for two replicates).

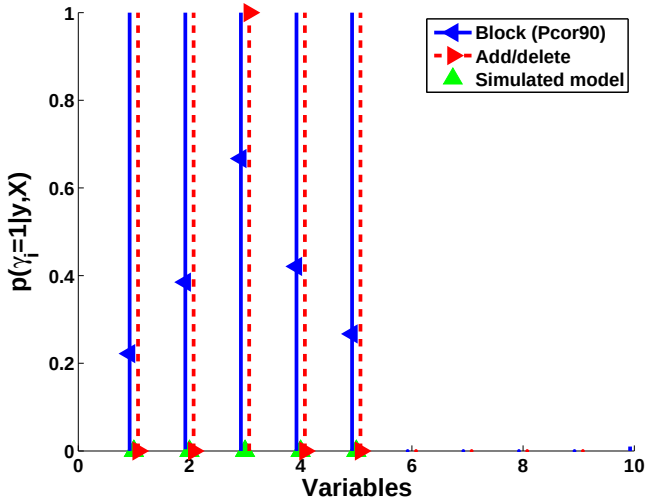


- $ESS^* = \frac{\#I}{p} \times \text{median}_{i \in I}(ESS(\gamma_i))$ where $I := \{i : ||\gamma_i|| > 0\}$
- Effective sample size $ESS(\gamma_i) = \frac{T-B}{f_i}$: Run length adjusted for integrated auto-correlation $f_i = 1 + 2\sum_{\kappa=1}^{\infty} \rho_\kappa(\gamma_i)$ (with $\rho_\kappa(\gamma_i)$ denoting auto-correlation of γ_i at lag κ)

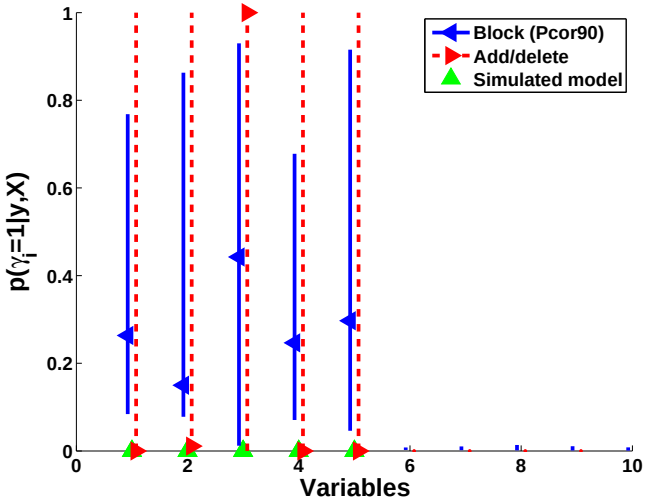
Simulation results

Posterior inclusion frequencies (median and IQR) for variables $1, \ldots, 10$ averaged over 25 replicates (after burn-in)

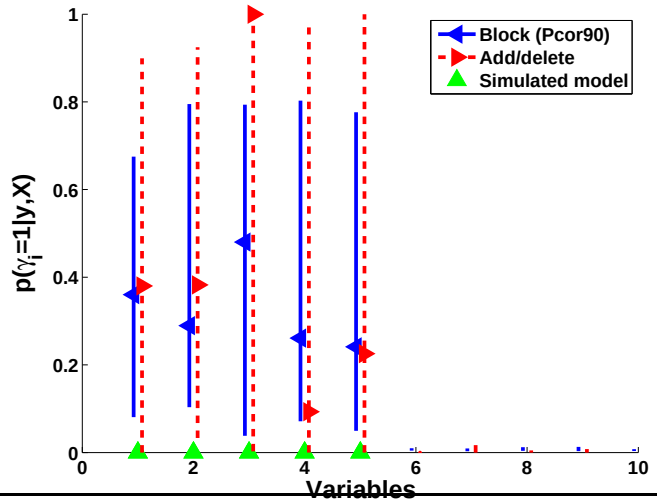
After 1,000 iterations



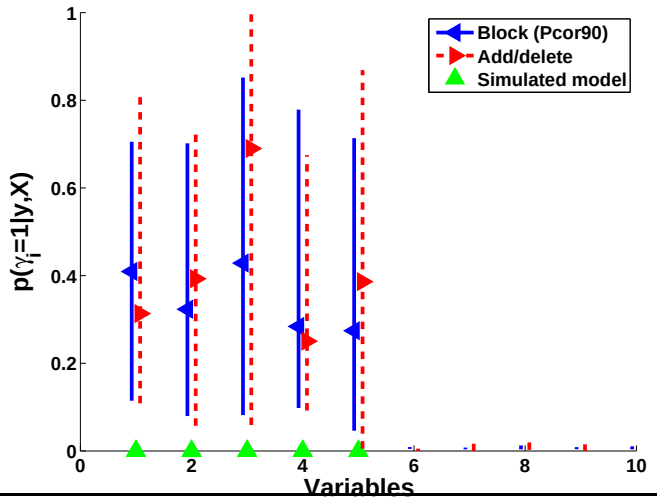
After 10,000 iterations



After 50,000 iterations



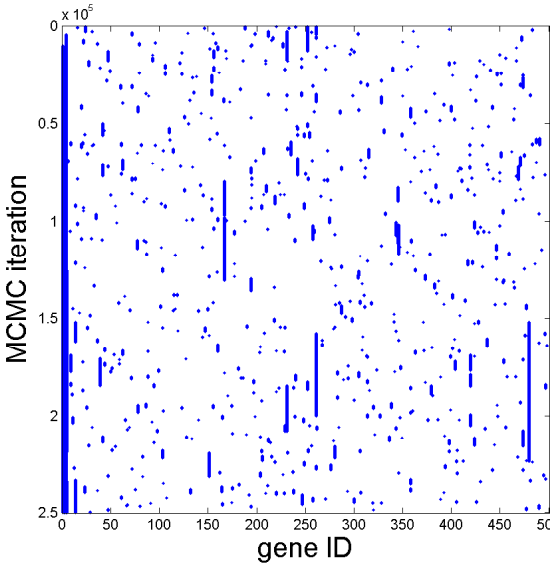
After 150,000 iterations



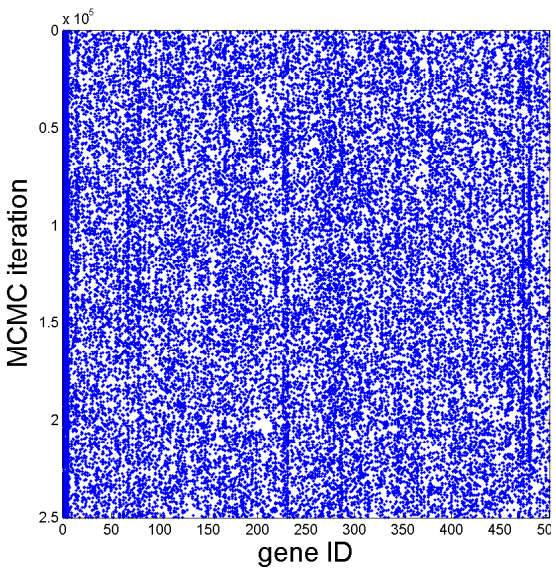
Simulation example B

Trace plots of γ vector show improved mixing for block update (results for one run)

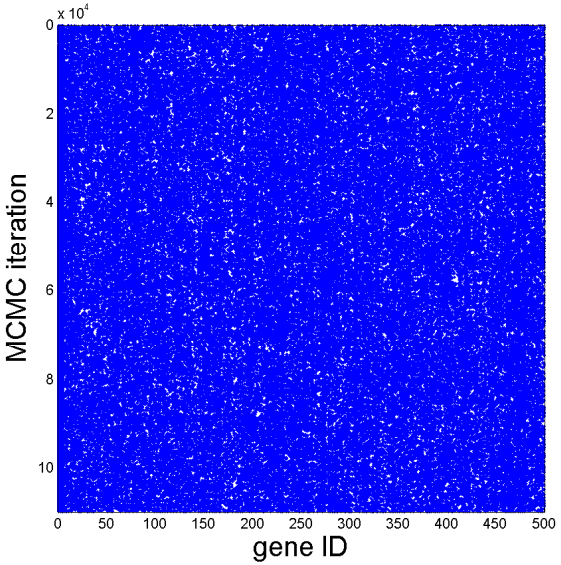
Add/delete sampler



Block (Pcor, $C = 90\%$)

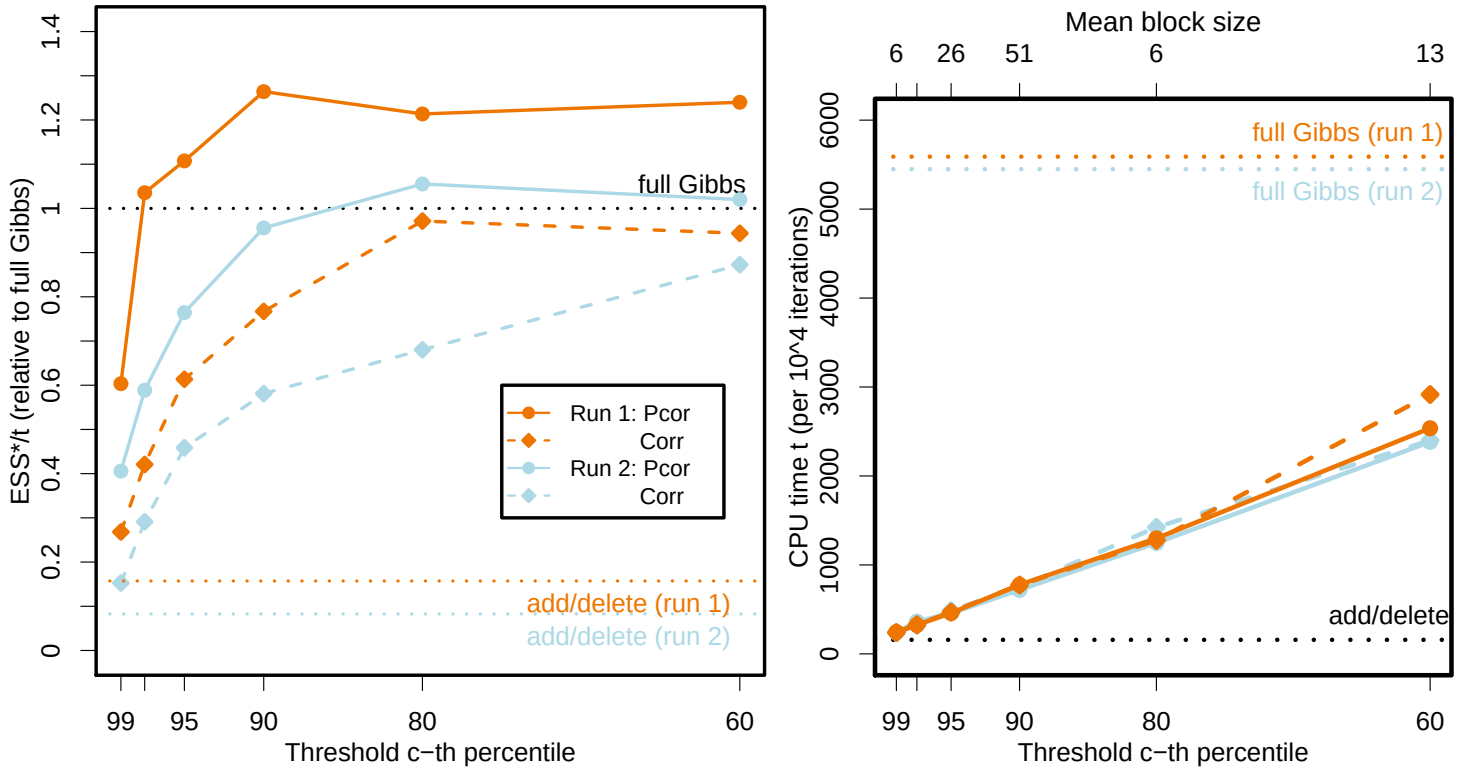


Full Gibbs



Simulation results

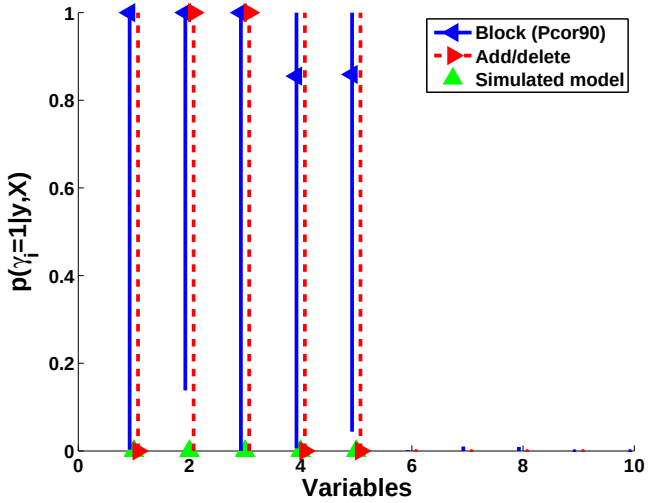
Relative median effective sample sizes (ESS^*/t) and CPU times t for various choices of block size threshold C (results for two replicates).



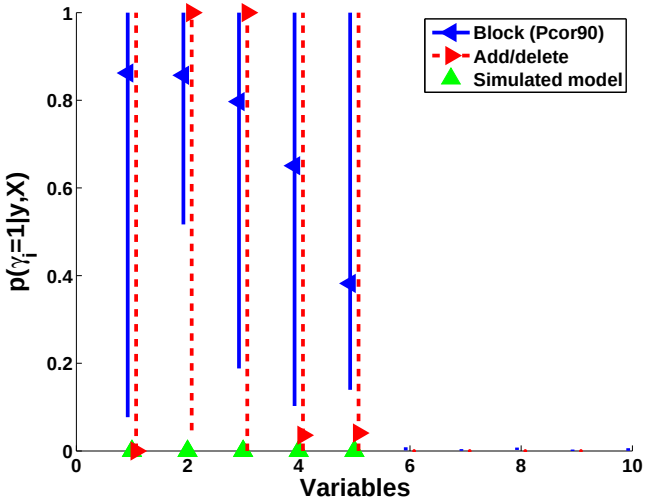
Simulation results

Posterior inclusion frequencies (median and IQR) for variables $1, \ldots, 10$ averaged over 25 replicates (after burn-in)

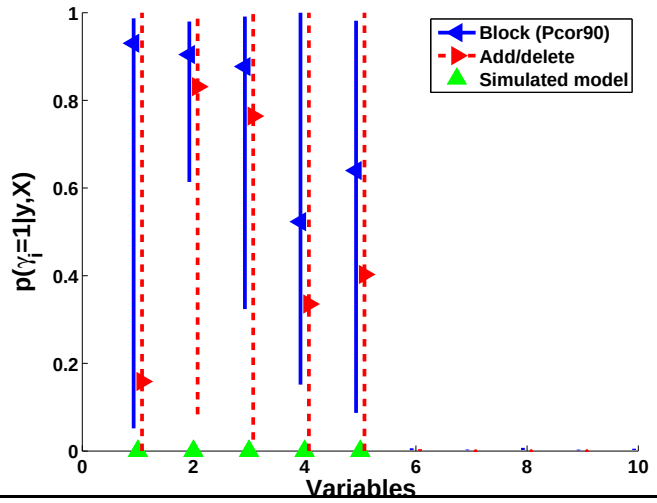
After 1,000 iterations



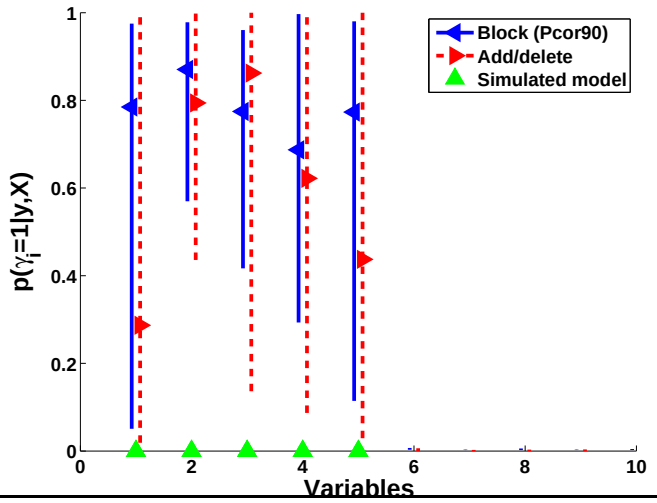
After 10,000 iterations



After 50,000 iterations



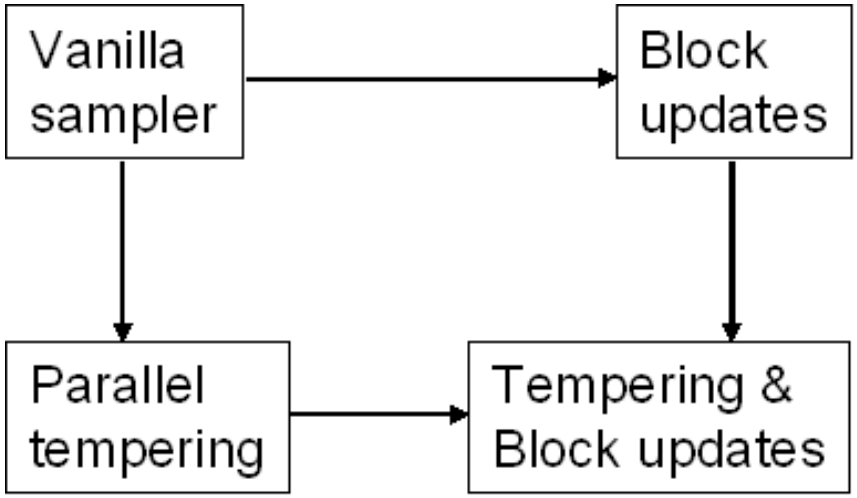
After 200,000 iterations



Parallel tempering on ovarian cancer data

Combine with other strategies to improve mixing

E.g. **Metropolis-coupled MCMC (parallel tempering):** Run parallel chains at different temperatures τ to improve mixing and propose swap



Parallel tempering implementation

- Tempered "likelihood" $\mathbf{p}_\tau(\mathbf{z}|\lambda, \beta_\gamma, \mathbf{X}_\gamma) = \mathbf{N}_\mathbf{z}(\mathbf{X}_\gamma\beta_\gamma, \tau\lambda)$
- Joint posterior distribution

$$\begin{aligned} p_\tau(\beta_\gamma, \gamma, z, \lambda|X_\gamma, y) &\propto p_\tau(\beta_\gamma, \gamma, z, \lambda, y|X_\gamma) \\ &= p(y|z)\mathbf{p}_\tau(\mathbf{z}|\lambda, \beta_\gamma, \mathbf{X}_\gamma)p(\beta_\gamma|\gamma)p(\gamma)p(\lambda) \end{aligned}$$

Swap the states of two parallel chains of temperatures τ_m and τ_l:

Exchange $(\beta_\gamma, \gamma, \lambda, z)$ with acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{p_m(z^{(l)}|\lambda^{(l)}, \beta_\gamma^{(l)}, X_\gamma^{(l)})p_l(z^{(m)}|\lambda^{(m)}, \beta_\gamma^{(m)}, X_\gamma^{(m)})}{p_m(z^{(m)}|\lambda^{(m)}, \beta_\gamma^{(m)}, X_\gamma^{(m)})p_l(z^{(l)}|\lambda^{(l)}, \beta_\gamma^{(l)}, X_\gamma^{(l)})}\right\} \\ &= \min\left\{1, \exp\left((\frac{1}{\tau_m} - \frac{1}{\tau_l})\left(-\frac{1}{2}(z_l - X_{\gamma l}\beta_{\gamma l})'\lambda_l^{-1}(z_l - X_{\gamma l}\beta_{\gamma l}) + \right.\right.\right. \\ &\qquad \left.\left.\left.\frac{1}{2}(z_m - X_{\gamma m}\beta_{\gamma m})'\lambda_m^{-1}(z_m - X_{\gamma m}\beta_{\gamma m})\right)\right)\right\} \end{aligned}$$

Parallel tempering on ovarian cancer data

Data set

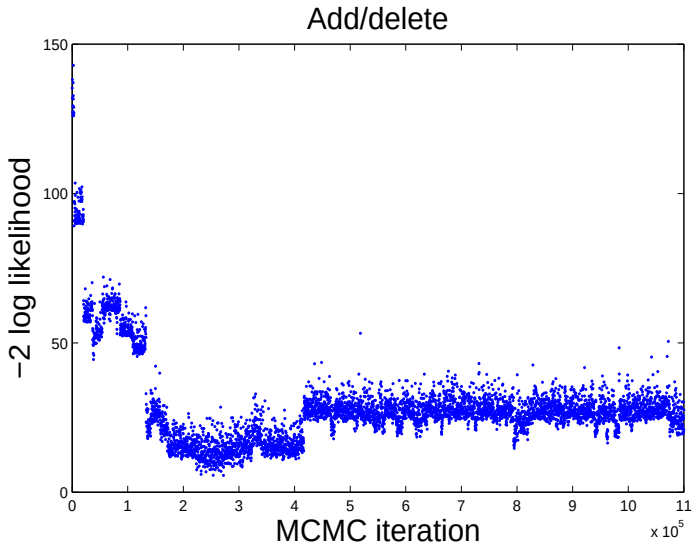
- Ovarian cancer gene expression data with $n = 104$ samples and $p = 4000$ variables (after univariate filtering) (Schwartz *et al.* 2002).
- Binary classification between intrinsically chemotherapy-resistant tumours and more responsive histologies.
- Block structure: partial correlation matrix and threshold $C = 0.99$
- In a previous resampling study, five genes had consistently been selected by lasso and other multivariate methods (for more than 50% of all training/validation splits) $\longrightarrow$ "candidate genes"
- Hyper-prior parameters in BVS model: $\pi = 5/4000$, $c^2 = 5$

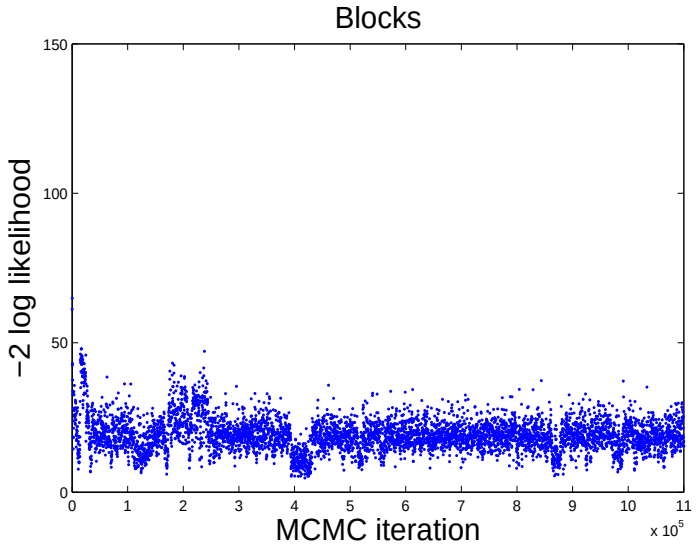
MCMC

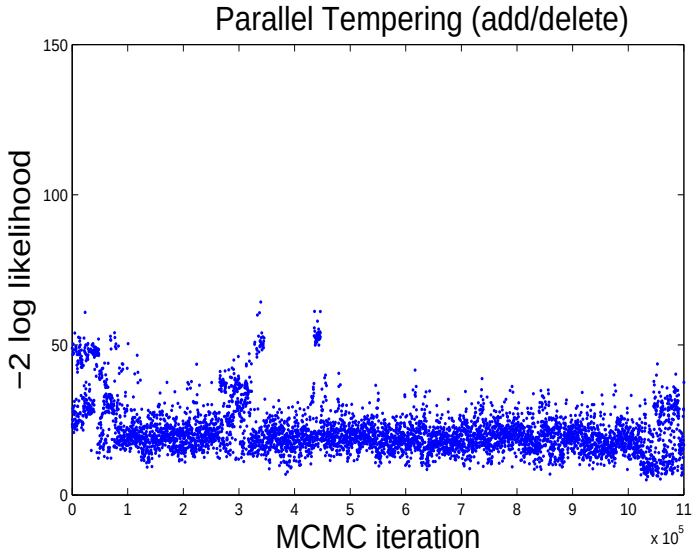
- $T = 1,100,000$, $B = 100,000$
- Five parallel chains with geometric temperature ladder $\{\tau^0, \tau^1, \tau^2, \tau^3, \tau^4\}$ with $\tau = 1.2$: Only propose to swap neighbouring chains.
  Run un-coupled for 50,000 iterations before starting swaps

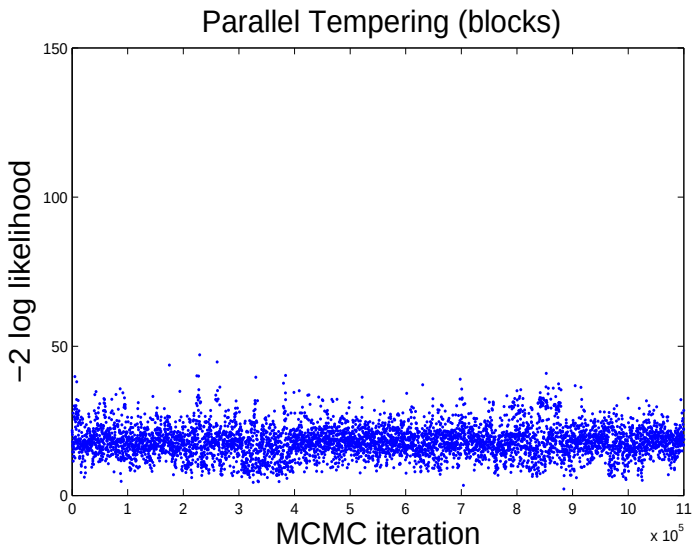
Parallel tempering on ovarian cancer data

Deviance trace shows convergence rate and when chain gets stuck in local optima.



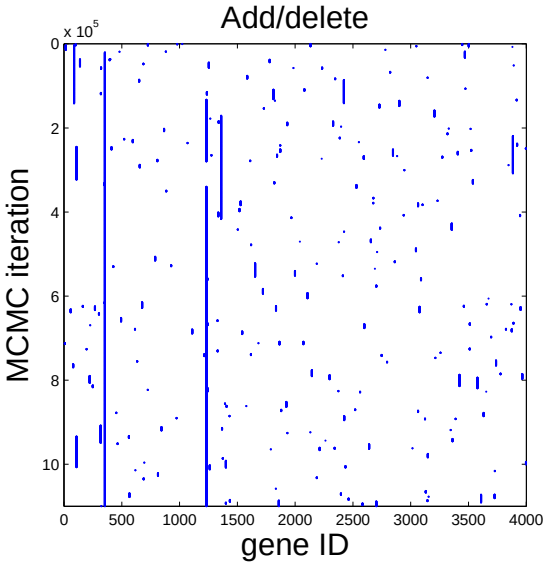


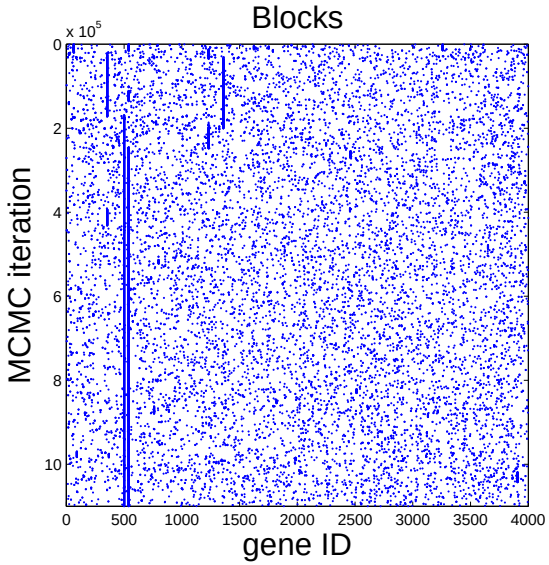


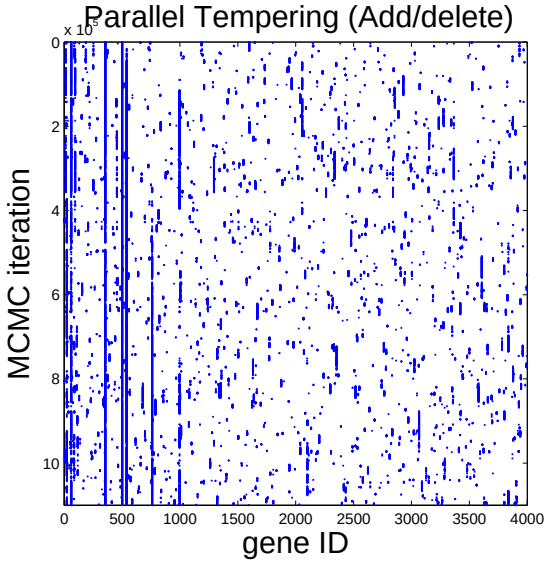


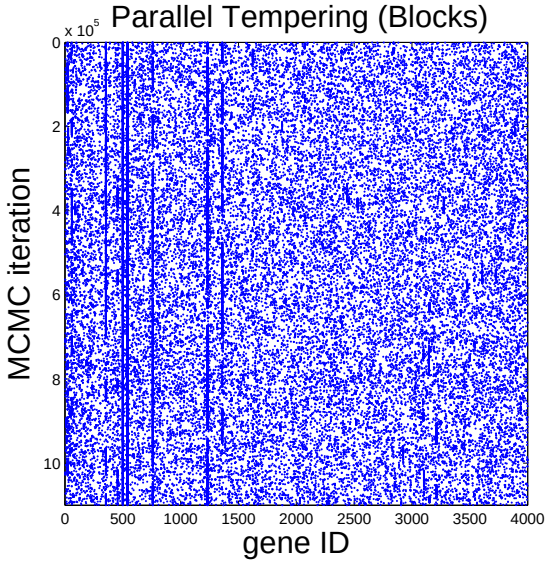
Parallel tempering on ovarian cancer data

Trace plots for γ: Mixing improves dramatically when using blocks/parallel chains.









Parallel tempering on ovarian cancer data

Diagnostic measures for mixing of Markov chains

MCMC sampler	CPU time t (sec)	ESS*	**ESS* $/t$ (to A/D)**	$\#I^\sharp$	# candidates†	# not candidates
Add/delete	18,906	8	**1.00**	**198**	1	23
Blocks	81,350	3,793	**114.4**	**2856**	3	6
Parallel tempering						
with add/delete	103,563	19,900	**471.4**	**1091**	4	15
with blocks	396,046	41,985	**260.1**	**3752**	4	5

$^\sharp I = \{i : ||\gamma_i|| > 0\}$

†How many of the five genes consistently selected by lasso and other multivariate methods in resampling study are recovered by BVS, if cut-off at **ratio of posterior to prior** > 10 (i.e. $P(\gamma_i = 1|Z, X, W) > 0.0125$)?

MCMC algorithms for variable selection

- Tempering with parallel chains is key to escape local modes
  - Important to allow bold moves, like “exchange moves”, but need to think carefully how to maximise their chance of being accepted (Evolutionary Monte Carlo, e.g. Bottolo and Richardson 2007, Jasra *et al.* 2005 etc.)
  - Ideal for exploiting parallel processing resources
- When there is dependence amongst the covariates, mixing can be improved by using block updating based on thresholding the (partial) correlations.
- We have explored many possible variants in terms of:
  - how to determine blocks (*Pcor*, *Corr*, random blocks)
  - size of blocks (choice of C)/ “block depth” (first-, second-,... order neighbours)
  - within block proposal: using a univariate Gibbs proposal, but for strongly correlated variables, could be useful to have additional moves proposing to update pair or triplets of variables using multivariate Gibbs

Conclusions

- Sensitivity analysis
  - choice of c^2 in prior covariance matrix $v = c^2 I_n$ of β
  - choice of π (prior prob $p(\gamma_i) = 1$)
- Choice of prior distributions:
  - for β_γ: independence prior, g-prior
  - for γ: binomial, beta-binomial
- Including some adaptivity is useful
  - Choice of temperature ladder
  - Investigate how to adapt number of models evaluated at each GS iteration
- Binary regression is harder to benchmark than linear regression. Not all variables are needed to obtain good 0/1 discrimination
  $\longrightarrow$ perform out of samples prediction to evaluate the algorithms
- Care is needed before interpreting results in the “large p, small n” case. Effective sample sizes are small!

Acknowledgements

Institute for Mathematical Sciences, Imperial College London

Leonardo Bottolo

Oxford Centre for Gene Function, Department of Statistics, University of Oxford

Chris Holmes

Financial support

wellcome trust