MCMC methods for Bayesian variable selection in binary classification in the "large p, small n" paradigm

Manuela Zucknick and Sylvia Richardson, **Department of Epidemiology and Public Health, Imperial College London (***Contact: manuela.zucknick03@imperial.ac.uk)*

1 Motivation: Improve slow MCMC mixing by updating covariates in blocks

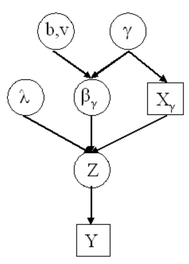
- In large-scale gene expression studies **vast numbers of genes** are scanned to find a small set of candidates which are linked to a particular phenotype.
- **Bayesian variable selection** (BVS) methods allow to solve such problems in a full probabilistic manner. Here, the Bayesian logistic regression approach by [2] is used for binary classification.
- **Markov chain Monte Carlo is used as a stochastic search algorithm** to find models with high posterior probability. The large scale of applications renders standard MCMC algorithms impractical (full Gibbs sampling too time-consuming, and fast single-variable addition/deletion algorithms mixing too slowly).
- Here, we **employ the dependence structure among covariates** to find variables to update together in blocks - to construct Markov chains which can move quickly around the vast model space.

2 Bayesian variable selection for logistic regression [2] and MCMC algorithm

Bayesian variable selection is implemented through a hierarchical model, where all possible 2^p models are represented by a p-dimensional indicator variable $\gamma_i = \begin{cases} 1 & \text{variable i is included} \\ 0 & \text{variable i is excluded} \end{cases}$

Bayesian hierarchical representation of logistic regression

$$
\begin{aligned}
y_j &= \begin{cases} 1 \text{ if } z_j > 0 \\ 0 \text{ otherwise} \end{cases} \\
z_j &= x_{\gamma j}\beta_\gamma + \epsilon_j \\
\epsilon_j &\sim N(0, \lambda_j) \\
\lambda_j &= (2\phi_j)^2 \\
\phi_j &\sim \text{Kolmogorov-Smirnov (i.i.d.)} \\
\beta_\gamma &\sim N(b_\gamma = 0, v_\gamma = c^2 \times I_{p_\gamma}) \\
\gamma &\sim p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i}.
\end{aligned}
$$



This hierarchical scale-mixture representation of the logistic regression model is useful because it preserves normal-normal conjugacy for z (see [2]).

Outline of MCMC algorithm

1. Update $\{z, \lambda\}$ jointly given $\{\beta_\gamma, \gamma\}$: $p(z, \lambda|\beta_\gamma, X_\gamma, y) = p(\lambda|z, \beta_\gamma, X_\gamma)p(z|\beta_\gamma, X_\gamma, y)$
2. Update $\{\beta_\gamma, \gamma\}$ jointly given $\{z, \lambda\}$ with joint proposal in a Metropolis-Hastings step

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

where the current covariate set (defined by γ) is updated, with a subsequent Gibbs update to β.

How to update the covariate set γ efficiently? The choice of the proposal distribution $q(\gamma^*)$ is crucial.

Possible proposal moves in $q(\gamma^*)$:

- **Add/delete move (*A/D*)**: select one γ_k at random and propose to change state (vanilla sampler)
- **Full Gibbs sampler (*Full*)**: updating all variables $i = 1, ..., p$ in each iteration by sampling from full conditional (not computationally feasible in large-scale problems)
- **Block samplers (*Block*)**: using the dependence structure to determine which covariates to update together. Select γ_k randomly, find neighbours nb and only for each γ_i in $nb \cup \gamma_k$ sample from full conditional (our approach).

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

1. Choice. **Correlation (*Corr*) or partial correlation (*Pcor*) matrix**: Estimate either using a shrinkage estimator (e.g. [3]).
2. Choice. **Threshold C**: Minimum size of absolute pairwise (partial) correlations for two variables to be declared neighbours. Here, the threshold is set to the c^{th} percentile of all absolute (partial) correlations (*Corr <c>, Pcor <c>*).

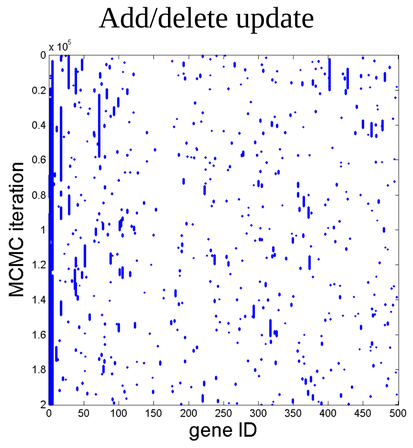
3 Simulation study (similar set-up to [1]): Large blocks outperform other MCMC methods

Repeat the following 25 times to create 25 simulated data sets:

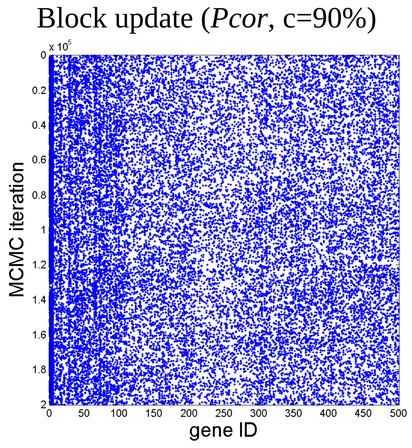
- $p = 100 \times 5$ variables and $n = 100$ samples, where $p^* = 5$ variables are related to response Y.
- 1. $X_1^*, ..., X_{100}^*$ iid $\sim N(0, 1)$
  2. $Z \sim N(0, 1)$; Simulate covariates as $X_i = X_i^* + Z$ $(i = 1, ..., 100) \rightarrow$ strong correlations
  3. Repeat step 2 five times to create $p = 100 \times 5$ variables
  4. Simulate Y (logistic model): $Y_j \sim \text{Bernoulli}\left(\frac{\exp(X_j\beta)}{1+\exp(X_j\beta)}\right)$ (j = 1,...,n), $\beta = (2, 2, 2, 2, 2, 0, ..., 0)$

Trace plots of γ vector show improved mixing for block update: (results for run 1)

Add/delete update



Block update (*Pcor*, c=90%)



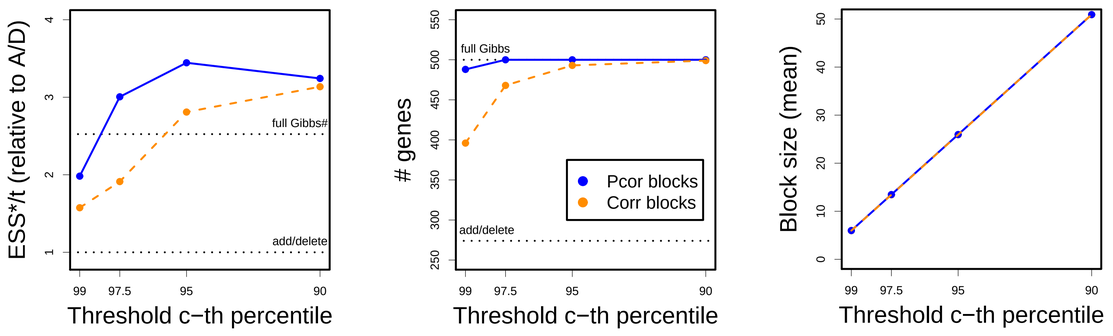
Diagnostic measures for mixing of Markov chains with respect to γ (200,000 MCMC iterations including 50,000 burn-in (except see ♯); hyper-prior parameters: $\pi = 5/500$, $c^2 = 5$)

Run 1:

Increase in ESS shows improved mixing. ESS^*/t relates ESS to computing time (t = time/10^4 sec)

With increasing block size more genes are considered.

Block sizes range from 4 to 50.



Runs 1 - 25 (averages): mean (range) of values	ESS^*/t (relative to A/D)	# genes†	Block size (mean)	# FP‡	# FN‡	# FP + # FN
add/delete update	1.00 (N/A)	270 (248,299)	N/A	12 (6,25)	1 (0,2)	13 (7,26)
block updates (*Pcor, c=90%*)	4.97 (2.65, 6.94)	500 (500,500)	50.89 (50.84,50.95)	8 (2,17)	1 (0,3)	9 (2,17)

Notes: Only variables included in at least one model are considered, ♯90,000 MCMC iterations including 10,000 burn-in samples
*Median effective sample size (per 10^4 iterations) median$(\text{ESS}_i) = \frac{T}{10^4 \tau_i}$:
sample size T divided by 10^4 iterations and by integrated auto-correlation $\tau_i = 1 + 2\sum_{\kappa=1}^{\infty}\rho_{i\kappa}$ for each γ_i
† number of variables for which $\gamma_i = 1$ in at least one MCMC iteration, ‡false positives and false negatives if cut-off at $P(\gamma_i = 1) > 0.05$

4 Application to gene expression data

- Ovarian cancer gene expression data with $n = 104$ samples and $p = 4000$ variables (after filtering) [4]. Binary classification between intrinsically chemotherapy-resistant tumours and more responsive histologies.
- Results are compared with our previous analysis using LASSO logistic regression [5], where a model with five genes was found to be the best discriminating model. One (*A/D*) or two (*Pcor, c=99.5%*) of these genes are recovered here (see Table)

Diagnostic measures for mixing of Markov chains with respect to γ
(550,000 MCMC iterations including 50,000 burn-in; hyper-prior parameters: $\pi = 5/4000$, $c^2 = 5$)

MCMC sampler	CPU time (sec)	ESS*	ESS inter-quartile range	ESS*/CPU time (relative to A/D)	# genes†	# neighbours (mean)	# genes in LASSO‡	# genes not in LASSO‡
add/delete update	15326	63.0	(28.7,128.3)	1.00	155	N/A	1	6
block update (*Pcor*, c=99.5%)	58516	2027.0	(708.4,6485.0)	8.43	1748	21.0	2	3

Notes: See previous Table. ‡How many of the five genes found by LASSO regression are recovered by the Bayesian variable selection model if cut-off at $P(\gamma_i = 1) > 0.05$, and how many are found in addition to the five LASSO genes.

5 Conclusions

- **Mixing is clearly improved by block updating** compared to the add/delete sampler. There is a trade-off for block size between improved mixing and increase in computation time reflected in the ratio ESS*/t. The **block size needs to be sufficiently large** to gain an improvement over the full Gibbs sampler.
- For this simulation set-up, the blocks constructed from **partial correlations perform slightly better** in terms of mixing and classification than blocks from correlations.
- We have investigated other ways than Gibbs sampling for updating γ within the blocks, in particular sampling from the **joint conditional distribution of all γ_i in the block**. However, the gain in mixing is not sufficient to off-set the increase in computation time.
- As the add/delete sampler mixes slowly, **long burn-in times are necessary**, which increase the computational burden. Here the same long burn-in time of 50,000 iterations was used for all samplers, which was sufficient even for A/D.
- Take care when interpreting results in "large p, small n": **Effective sample sizes are small!**

Acknowledgements We thank Chris Holmes for many fruitful discussions. MZ receives financial support for her PhD by the Wellcome Trust.

References

[1] E.G. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[2] C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006.

[3] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32, 2005.

[4] D.R. Schwartz *et al.* Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. *Cancer Research*, 62:4722–4729, 2002.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.