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

M. Zucknick[†*], S. Richardson[*]

[*] *Centre for Biostatistics, Department of Epidemiology and Public Health, Imperial College London*

[†] E-mail: *manuela.zucknick@imperial.ac.uk*

In large-scale genomic applications vast numbers of markers or genes are scanned in order to find a small number of candidates which are linked to a particular disease or phenotype. This is a variable selection problem in the "large p, small n" paradigm where many more variables than samples are available. Additionally, a complex dependence structure is often observed among the markers/genes due to their joint involvement in biological processes and pathways.

Bayesian variable selection methods which use an indicator variable to signify which variables are considered included in the model at every iteration are well suited to the problem. If the phenotype is binary the variable selection problem is cast in the context of binary classification. Here we follow the implementation of the Bayesian logistic regression model by Holmes and Held (2006) which leads to marginal Gaussian distributions.

However, because of the vastness of the model space, full posterior inference by Markov chain Monte Carlo is not feasible and MCMC methods are used instead as stochastic search algorithms with the aim to quickly find regions of high posterior probability. Simple addition/deletion/swap Metropolis-Hastings proposals are fast to evaluate which is the main reason why they have been applied in this context (e.g. Sha et al. 2004). However, it has been noted that such proposals experience problems if $p >> n$ in that the acceptance probability for deleting variables tends to zero (Hans et al. 2007). Full Gibbs sampling on all variables, on the other hand, is computationally very expensive for large p. So instead of updating every variable in every iteration, we propose to utilise the dependence structure among the genes/markers in order to decide which variables to update in a block at each iteration.

The mixing and convergence performances of the resulting Markov chains are evaluated and compared to standard samplers in both a simulation study and in an application to a real gene expression data set.

References

C. Hans, A. Dobra, and M. West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 2007. (to appear).

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

N. Sha, M. Vannucci, M.G. Tadesse, P.J. Brown, I. Dragoni, N. Davies, T.C. Roberts, A. Contestabile, N. Salmon, C. Buckley, and F. Falciani. Bayesian variable selection in multinomial probit models to identify molecular singatures of disease stage. *Biometrics*, 60:812–819, 2004.