MCMC methods for gene expression profiling via Bayesian variable selection

Manuela Zucknick[1,2] and Sylvia Richardson[2]

[1] DKFZ, Im Neuenheimer Feld 280, D-69120 Heidelberg m.zucknick@dkfz.de
[2] Biostatistics Centre, Imperial College, Norfolk Place, London W2 1PG, UK

Abstract. Gene expression microarrays and other high-throughput technologies produce measurements for several thousand genes but the sample size is typically much smaller than that. A common application of microarray data is the construction of gene expression profiles for class prediction based on expression measurements of a small number of selected genes.

Bayesian variable selection methods are well suited to the problem, since priors can be imposed so that full modelling is possible even if the number of variables is much larger than the sample size. In addition, the uncertainty related to the role of each candidate gene can be assessed through posterior probabilities. However, the model space is very large and standard MCMC algorithms are computationally unfeasible. Several strategies can be used and ultimately combined to improve feasibility. Here, we focus on the 'block update' component of an MCMC strategy. We propose to employ the dependence structure in the data to decide which variables should always be updated together and which are nearly conditionally independent and do not need to be considered together.

For binary classification, logistic regression is traditionally preferred in medical and biological applications because of the good interpretability. We follow the implementation of the Bayesian logistic regression model by Holmes and Held (2006).

We investigate several MCMC samplers using the dependence structure in different ways. The mixing and convergence performances of the resulting Markov chains are evaluated and compared to standard samplers using simulated data and in an application to a gene expression data set related to ovarian cancer. In the latter, we also explore the additional benefit of combining the block update with a parallel tempering strategy.

References

HOLMES, C.C. and HELD, L. (2006): Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis, 1, 55–67.*

Keywords

GENE EXPRESSION, CLASSIFICATION, VARIABLE SELECTION, MCMC