Bayesian sparse regression models of high dimensional data

Leonardo Bottolo and Sylvia Richardson*

Centre for Biostatistics, Imperial College London

1 Introduction

In parallel to fast evolving biotechnology that give rise to high dimensional data in many set-ups, there is increased interest in searching for sparse structure in such high dimensional data sets. In particular, multivariate regression models with a very large number of predictors together with multiple responses have attracted the attention of the statistical community in very recent years, as numerous case studies are designed so as to be able to link genetics information to a wide range of -omics measurements in order to advance functional knowledge. A notable example is the paradigm of eQTL analysis, where thousands of transcripts measurements are regressed versus (hundred of) thousands of markers. In this context the usual problem of multimodality of the posterior distribution, when $p >> n$, is further complicated by the dimension of the response matrix. In this large p, small n framework, a sparse representation is typically imposed, i.e. parsimonious models containing only a few predictors are sought to gain interpretability.

2 Models and Methods

In this talk, we discuss Bayesian variable selection in the general context of linear regression models with multiple outcomes, highlighting the key prior choices that induce sparsity. We introduce a new searching algorithm called Evolutionary Stochastic Search (ESS) (Bottolo and Richardson, 2009) which helps to find sets of covariates that predicts responses' variation. ESS is based on ideas from population Monte Carlo, and entails runnning multiple chains at different temperature, with exchange of information between the chains. We have used ESS as a building block for two different classes of models:

(i) When the number of outcomes q is not too large with respect to the number of observations $n, n < q$, extension of the linear regression model to model simultaneously multiple outcomes, Sparse Bayesian Multiple Regression (SBMR), is applied using ESS as search engine. To reduce the computational burden, most of the regression parameters are integrated out, leaving only the variable selection indicators and key variance parameters to update.

(ii) When the number of outcomes is large (in particular, larger than the number of observations), a new model, Hierarchical Evolutionary Stochastic Search, HESS, which links the q linear regression models and their sparse formulation in a hierarchical way, is introduced. In this model, a version of ESS that has been designed to sample efficiently from the vast parametric space is used to obtain posterior samples of allocations. Moreover, inspired by adaptation techniques (Roberts and Rosenthal, 2006), we have implemented versions of the HESS algorithm that continuously adapt to focus exploration on a subset of q responses where the model choice is more critical. The benefits of setting up suitable adaptive shemes will be illustrated; this is work in progress.

3 Applications

We have analysed simulated and real data sets to demonstrate the performance of the proposed algorithms. For the SBMR model, we will present the results obtained analysing jointly gene expression levels for several tissues in an animal data model. Here the model is used to test the biological hypothesis of pleiotropy, i.e. shared genetic regulators for the gene expression across tissues. We will show how it provided a powerful alternative to intersecting results from univariate analyses in each tissue.

Secondly, we will discuss the performance of HESS in a challenging example of human metabolomic data where common genetics regulators of a 300 bins discredited Mass Spectography are sought. In both real data examples, the number of covariates (markers and suitable chosen SNPs from a Genome Wide Scan), 770 and 5,000 respectively, largely outnumbers the observations, 29 and 50 respectively, representing challenging analyses.

References

Bottolo, L. and Richardson, S. (2009) Evolutionary stochastic search. *Journal of Computational and Graphical Statistics*, under revision.
Available at http://www.bgx.org.uk/publications.html

Roberts, G.O. and Rosenthal, J.S. (2006) Examples of Adaptive MCMC *J. Comp. Graph. Stat.* to appear.