ATELIER « MÉTHODES STATISTIQUES COMPUTATIONNELLES EN GÉNOMIQUE ET EN BIOLOGIE SYSTÉMIQUE »

18–22 AVRIL 2011

WORKSHOP ON "COMPUTATIONAL STATISTICAL METHODS FOR GENOMICS AND SYSTEMS BIOLOGY"

APRIL 18–22, 2011

Sparse Bayesian regression in genomics

Sylvia Richardson *

sylvia.richardson@imperial.ac.uk

In genetics and genomics, one of the most sought after statistical output is a set of reproducible sparse regression models, i.e. models that select only a few relevant predictors (i.e. risk factors, SNPs, transcripts) amongst a very large set of possible candidates, together with good assessment of how uncertain their role is.

In this talk, we shall focus on the task of building efficient regression models for sparse multivariate analysis of high dimensional data sets, starting from models that analyse a small number of phenotypes, as is typical in GWAS association studies, and extending to cases where the numbers of responses q, and of predictors p, to analyse jointly are both large with respect to the sample size n, a common case in genetical genomics studies.

We shall review first the Bayesian variable selection set-up for the linear model and an efficient MCMC algorithm, ESS++, that has been developed to search high dimensional model spaces (Bottolo and Richardson, 2010, Bottolo *et al*, 2011). We will illustrate how ESS++ can be used in GWAS studies by taking advantage of cutting-edge linear algebra libraries that employ GPU technologies. We will also discuss alternative penalised regression approaches, notably the implementation of Stability Selection approaches (Meinshausen and Bühlman, 2010) in the GWAS context.

In a second part, we will discuss the extended set-up of hierarchical related sparse regressions, where parallel regressions of a large set of responses on the same set of covariates are linked in a hierarchical fashion, a challenging bi-directional situation of the 'large p, small n' paradigm and a common set up in many integrative genomics analyses. We will outline prior models for the variable selection indicators, which correspond to compromises between the aims of controlling sparsity and that of borrowing information across the responses to enhance discovery of key predictors, also referred to as "hot spots". We will outline recent developments of adaptive MCMC algorithms that may be suitable in such set ups. They permit progressively focussing on part of the space where 'the action' happens. Models and algorithms will be illustrated on simulated and real data from genomics. The talk will conclude by a brief discussion of future issues and challenges in this field.

Joint work with Leonardo Bottolo and David Hastie (*Imperial College*) *and Ismail Ahmed* (*INSERM*).

*School of Public Health, Imperial College London, South Kensington Campus, London, SW7 2AZ, United Kingdom.