BIG DATA IN BIOMEDICINE. BIG MODELS?

WORKSHOP PROGRAM

Session 1 (9:30 - 11:30). Room LIB1, Library building

9:30 **David Rossell** (University of Warwick, UK)
Welcome address

9:40 **Keynote: Valen Johnson** (Texas A&M University, USA).

Title: *Implications of uniformly most powerful Bayesian tests for the reproducibility of scientific research*

Abstract: Connections between uniformly most powerful Bayesian tests and classical uniformly most powerful tests are used to equate the rejection regions of these two types of tests in order obtain an approximate equivalence between Bayes factors and p-values. The implications of this equivalence for the reproducibility of scientific research are examined. Approximately uniformly most powerful Bayesian tests are described for t tests, and the power of these tests are compared to ideal Bayes factors (defined by determining the best test alternative for each true value of the parameter), as well as to Bayes factors obtained using the true parameter value as the alternative. Interpretations and asymptotic properties of these Bayes factors are also discussed.

10:20 **Rich Savage** (University of Warwick, UK)

Title: *The data deluge and what to do with it*

Abstract: Medicine has become a data-rich subject. This opens up huge possibilities for diagnostic/prognostic prediction as a tool for doctors to help patients. But to do this, we must develop ways to use effectively a hugely diverse range of data sources. I'll discuss the current state-of-the-art in medical bioinformatics and take a look at some of the future challenges that we're about to face.

10:50 **Jun Liu** (Harvard University, USA)

Title: *Bayesian inverse multivariate regression with application to eQTL analysis*

Abstract: Expression quantitative trait loci (eQTLs) are genomic locations associated with changes of expression levels of certain genes. By assaying gene expressions and genetic variations simultaneously on a genome-wide scale, scientists wish to discover genomic loci responsible for expression variations of a set of genes. The task can be viewed as a multivariate regression problem with variable selection on both responses (gene expression) and covariates (genetic variations), including also multi-way interactions among correlated covariates. Instead of learning a predictive model of quantitative trait given combinations of genetic markers, we adopt an inverse modeling perspective to model the distribution of genetic markers conditional on certain quantitative traits. A particular strength of our method is its ability to detect interactive effects of genetic variations with high power even when their marginal effects are weak, addressing a key weakness of

many existing eQTL mapping methods. Furthermore, we illustrate how the methodology can be adapted to the general semi-parametric regression framework based on index modeling, and how to design a non-iterative step-wise algorithm to discover multi-way interactions among predictors in O(p) time, with p being the number of predictors in consideration. This is based on the joint work with Bo Jiang.

11:20 **Floor discussion**

Break (11:30 - 11:50)

Session 2 (11:50 - 13:00). Room LIB1, Library building

11:50 **Mark Girolami** (University of Warwick, UK)

Title: *Bayesian Model Selection: An Aid to the Cell Biologist?*

Abstract: Understanding the mechanisms of cell signal transduction holds the promise of new targeted cancer therapies as suggested by the success of specific Tyrosine Kinase Inhibitors. Formal models of biochemical kinetics describing hypothesised components of signalling pathway structures provide a means of assessing their evidential support by data assimilation. Bayesian model selection is a conceptually simple and appealing framework with which to evidentially assess competing hypothesised structures and this talk will discuss how it may aid the scientific process in cellular biology. In addition to considering a number of conceptual, methodological and technical issues that arise two ongoing case studies, being undertaken with cancer biologists, based on (1) an investigation into the regulation of the EWS-FL1 translocation fusion protein associated with Ewings Sarcoma, and (2) the interplay of RAF isoforms in the MAPK pathway are presented.

12:20 **Peter Mueller** (UT Austin, USA)

Title: *A Bayesian Feature Allocation Model for Tumor Heterogeneity*

Abstract: We characterize tumor variability by hypothetical latent cell types that are defined by the presence of some subset of recorded SNV's. (single nucleotide variants, that is, point mutations). Assuming that each sample is composed of some sample-specific proportions of these cell types we can then fit the observed proportions of SNV's for each sample. In other words, by fitting the observed proportions of SNV's in each sample we impute latent underlying cell types, essentially by a deconvolution of the observed proportions as a weighted average of binary indicators that define cell types by the presence or absence of different SNV's. Taking a Bayesian perspective, we proceed with a prior probability model for all relevant unknown quantities, including in particular a prior probability model on the binary indicators that characterize the latent cell types by selecting (or not) the recorded SNV's. Such prior models are known as feature allocation models. We define a simplified version of the Indian buffet process, one of the most traditional feature allocation models.

12:50 **Floor discussion**

Lunch break (13:00 - 14:00)

Session 3 (14:00 - 15:50). Room L5, Science Concourse

14:00 **Keynote. Aki Vehtari** (Aalto University, Finland)

Title: On Bayesian predictive methods for high-dimensional covariate selection

Abstract: In this talk, I discuss the bad behavior of commonly used Bayesian model selection methods, such as the deviance information criterion (DIC), the widely applicable information criterion (WAIC), Bayesian cross-validation, relative posterior probabilities, and Bayes factors, in high-dimensional covariate selection. Even if the model selection is based on an approach giving unbiased expected predictive performance estimates for any particular model, the data-fitted model selection procedure causes the expected predictive performance estimate of the selected model to be biased. If the number of candidate models is very large, such a model selection procedure can strongly overfit to the data. The same effect can be observed when relative posterior probabilities or Bayes factors are used for the model selection, which is not surprising since the marginal likelihood can be interpreted as a sequential predictive criterion. I also present an alternative decision theoretical solution which can avoid the model selection induced bias and overfitting. I review model selection methods which Vehtari & Ojanen (2012) call reference predictive methods, and discuss computational challenges in their use.

14:40 **Donatello Telesca** (UCLA, USA)

Title: *Mixture representations of Non Local priors and graphical model determination and estimation*

Abstract: We review the general representation of non local priors as mixtures of truncated distributions. This construction is applied to the definition of non local priors for selection parameters of directed and undirected Gaussian graphical models. Several constructive definitions are explored and related posterior simulation strategies compared. We focus our investigations on parameter estimation in a comparative study.

15:10 **Guido Consonni** (Universita Cattolica di Milano, Italy)

Title: *Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors*

Abstract: Directed Acyclic Graphical (DAG) models are increasingly employed in the study of physical and biological systems to model direct influences between variables. Identifying the graph from data is a challenging endeavour, which can be more reasonably tackled if the variables are assumed to satisfy a given ordering; in this case we simply have to estimate the presence or absence of each potential edge.Working under this assumption, we propose an objective Bayesian method for searching the space of Gaussian DAG models, which provides a rich output from minimal input. We base our analysis on non-local parameter priors, which are especially suited for learning sparse graphs, because they allow a faster learning rate, relative to ordinary local parameter priors, when the true unknown sampling distribution belongs to a simple model. We implement an efficient stochastic search algorithm, which deals effectively with data sets having sample size smaller than the number of variables, and apply our method to a variety of simulated and real data sets. Our approach compares favourably, in terms of the ROC curve for edge hit rate versus false alarm rate, to current state-of-the-art frequentist methods relying on the assumption of ordered variables; under this assumption it exhibits a competitive

advantage over the PC-algorithm, which can be considered as a frequentist benchmark for unordered variables. Importantly, we find that our method is still at an advantage, for learning the skeleton of the DAG, when the ordering of the variables is only moderately mis-specified. Prospectively, our method could be coupled with a strategy to learn the order of the variables, thus dropping the known ordering assumption.

15:40 **Floor discussion**

Break (15:50 - 16:10)

Session 4 (16:05 - 17:15). Room L5, Science Concourse

16:05 **Jianhua Hu** (MD Anderson Cancer Center, USA)

Title: *High dimensional variable selection for correlated data*

Abstract: Correlated data are usually encountered in high throughput genomic and proteomic data. We propose a Bayesian model selection procedure to identify important variable clusters (e.g., gene sets) that are predictive of a disease phenotype. The clustering algorithm appears to have close connections with several existing clustering algorithms including K-means. We adopt the product moment prior on the variable parameter space to control for false positive rate in the high dimensional problem. The proposed procedure is shown to have appealing empirical performance include the case of p¿n.

16:35 **Luca La Rocca** (Universita di Modena e Reggio Emilia, Italy)
Title: *Cutting-edge issues in objective Bayesian model comparison*
Abstract: When a Bayes factor is used to compare two nested models, the parameter priors under the two models determine its performance as an implementation of Ockhams razor (the principle that an explanation should not be more complicated than necessary). On the one hand, asymptotically, the rate at which evidence accumulates in favour of the smaller model radically depends on the local or non-local character of the prior under the larger model. On the other hand, for small samples, the amount of evidence in favour of the smaller model will be critically inflated by diffuseness of the prior under the larger model. Moreover, parameter priors can make a difference in terms of finite-sample ability to drop the larger model when both models perfectly fit the data. In my talk, based on some recent work dealing with these issues, I shall discuss, from an objective standpoint, how sharp an edge should Bayesian barbers put to their tools, assuming they care about not cutting the throat of the larger model.

17:05 **Floor discussion**