different methods and populations.

A30
Multiple Kernel Learning for Genomic Predictions of Complex Traits

Athina Spiliopoulou, Dominik Glodzik, Mairead Bermingham, Caroline Hayward, Igor Rudan, Harry Campbell, Alan Wright, Ricardo Pong-Wong[2], Pau Navarro, Chris Haley, Felix Agakov

MRC Human Genetics Unit, University of Edinburgh
[2]The Roslin Institute, University of Edinburgh

Recent years have seen great technological advances in the collection of high quality '-omics' data and significant progress in analysing and associating it with biological function. Here our goal is to leverage the abundance of available information into computational models that can accurately predict complex phenotypic traits in humans. To do this, we build on advances in non-parametric methods, where predictions for new individuals are made according to their similarity to other individuals from training cohorts, with the similarity *(kernel)* functions learned from data. Kernel methods are particularly suited for performing inferences with biological data, as they can be defined for a wide range of data types including vectors, strings and graphs. These kernel constructions correspond to different notions of similarity and often use complementary sources of information. Therefore, considering a combination of kernels is potentially more powerful than any single one on its own. Recent advances in the machine learning community have led to Multiple Kernel Learning (MKL) algorithms that can automatically discover the most appropriate kernel combination for a prediction task, typically by applying l1 or l2 type penalties on the weights of the kernel combination.

In this work we first apply a number of standard kernels from the machine learning literature, and string kernels based on either a haplotype or a genotype representation of the data. We then construct novel kernels that utilise the wide range of available genomic annotations, such as GWAS meta-analysis hits and eQTLs, or exploit domain knowledge by slicing the genome into new and old segregating variants. We present results from carefully designed cross-validation experiments that evaluate the performance of the MKL framework on predicting height, body mass index (BMI) and high density lipoproteins (HDL) in a Croatian population cohort. We examine the merits of using multiple versus single kernels and perform a comparative analysis with three parametric models commonly used in the literature, namely ridge regression, lasso and the elastic net.

A31
A novel two stage approach for causal effect estimation based on predictions of the response variable

Desmond Dedalus Campbell, Sabine Landau[2], Pak Chung Sham, Jo Knight[3]

University of Hong Kong
[2]Institute of Psychiatry, Kings College London
[3]Centre of Addiction and Mental Health

Research builds upon pre-existing knowledge by investigating the relationship between novel variables and variables whose inter-relationships are well understood. This scenario can often be described as a complex model of known structure incorporating latent and manifest variables. In parts of the model relating to well understood variables, parameter values might be well estimated as a result of previous research. For the rest of the model the parameters are unknown. The research question can be framed as the estimation of these unknown model parameters. Traditionally this would be addressed by fitting to the data a Structural Equation Model (SEM) incorporating well understood and novel manifest variables.
We propose an alternative two stage analysis approach for causal parameter estimation. First various evidence (including observations on the well understood variables) is brought to bear via a linear SEM on predicting values for a latent variable representing the trait of interest. The second stage is then a regression of the latent variable predictions on candidate explanatory variables.
Very large computational savings may result when the first stage (done only once) is computationally expensive, while its output is reused in many computationally cheap second stages. This makes feasible the use of complex trait models that would be prohibitively expensive to fit via the traditional approach. For instance, one scenario that might benefit from our two stage approach is where liability in disease pedigree members (predicted using Markov Chain Monte Carlo methods) is regressed on their GWAS chip genotypes.
Our two stage approach also allows the analysis to be partitioned; latent variable distributions can be generated by specialist modellers, while novel variable effects can be assessed by non-specialists. It facilitates the exploitation of previous research findings and increases comprehensibility by summarising evidence in predicted latent variable distributions.
We present a novel method for the second stage of our two stage approach and compare it to a more orthodox method via simulation studies.

A32
Fast mixed models for whole-genome association studies

Yurii Aulchenko

Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia; YuriiA consulting, The Netherlands

In this talk, I will present current state of an active research area - development and application of mixed models for analysis of correlated data in the context of whole-genome association studies. I will review and outline the relations between different classes of approximations to full Maximum Likelihood-based mixed models, and discuss several recent advances in the field. Both statistical and computational aspects of the problem of genome-wide association analysis of