

CSI Special One-Day Meeting, 26 September 2011, Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

Titles and abstracts of talks

Kees Albers, Wellcome Trust Sanger Institute and Dept of Haematology

Estimating statistical significance of exome sequencing data for rare mendelian disorders using population-wide linkage analysis

Exome sequencing of a small number of unrelated affected individuals has proved to be a highly effective approach for identifying causative genes of rare mendelian diseases. A widely used strategy is to consider as candidate causative mutations only those variants that have not been seen previously in other individuals, and those variants predicted to affect protein sequence, e.g. non-synonymous variants or stop-codons.

For the recessive disorder Gray Platelet Syndrome we identified 7 novel coding mutations in 4 affected individuals, all in different locations in one gene and absent from 994 individuals from the 1000 Genomes project; intuitively a highly significant result (Albers et al. Nat Genet 2011). However, in the case where the candidate causative mutations segregate at low frequency in the general population the significance may be less obvious. This raises a number of questions: what is the statistical significance of such findings in small numbers of affected individuals? If we would assume that the causative mutations are not necessarily in coding sequence, would these results be genome-wide significant? Motivated by these issues, we are developing a statistical model based on the idea that filtering out previously seen variants can be thought of as performing a whole-population parametric linkage analysis, whereby the individuals carrying previously seen variants represent the unaffected individuals.

We use the coalescent, a mathematical description of the notion that ultimately all individuals in a population are descendants of a single common ancestor, to model the unknown pedigree shared by the affected individuals and the unaffected individuals.

I will discuss implications of population stratification, false positive variant calls and variation in coverage for singleton rates and significance estimates.

Rosemary Bailey, School of Mathematical Sciences, Queen Mary, University of London

Design and analysis of biodiversity experiments

Colleagues in ecology designed an experiment to see whether various favourable responses were affected by the number of different species present in the ecosystem, keeping the total number of organisms constant.

I thought that their data were better explained by a model that was more obvious to me. I will describe the experiment, the family of models we discussed, the conclusion from the data analysis, and the design of subsequent studies.

Andrei Bejan, Statistical Laboratory

Using velocity fields in evaluating urban traffic congestion via sparse public transport data and crowdsourced maps

It is widely recognised that congestion in urban areas causes financial loss to business and increased use of energy compared with freeflowing traffic. Providing one with accurate information on traffic conditions can encourage journeys at times of low congestion and uptake of public transport. Installing a static measurement infrastructure in a city to provide this information may be an expensive option and potentially invade privacy. Increasingly, public transport vehicles are equipped with sensors to provide realtime arrival time estimates, but these data are fleet specific and sparse. The recent work with colleagues from the Cambridge University Computer Laboratory showed how to overcome data mining issues and use this kind of data to statistically analyse journey times experienced by road users generally (i.e. journey durations experienced by public transport users as well as individual car drivers) and influence of various factors (e.g. time of day, school/out of school term effects, etc)[Be10, Ba11]. Furthermore, we showed how the specifics of these location data may be used in conjunction with other sources of data, such as crowdsourced maps, in order to recover speed information from the sparse movement data and reconstruct information on transport traffic flow dynamics in terms of velocity fields on road networks[Be11]. In my short talk I will present a number of snapshots illustrating this analysis and some results and introduce the problem of comparing/classifying velocity fields and early spotting of accidents and their consequences for the traffic and road users.

References

- [Ba11] Bacon, J., Bejan, A., Evans, D., Gibbens, R., Moody, K. (2011) Using Real-Time Road Traffic Data to Evaluate Congestion. To appear in *Lecture Notes in Computer Science*, 6875. Springer.
- [Be11] Bejan A., Gibbens R. (2011) Evaluation of Velocity Fields via Sparse Bus Probe Data in Urban Areas. To appear in *14th International IEEE Conference on Intelligent Transportation Systems*, Washington DC, USA, October 5-7.
- [Be10] Bejan A., Gibbens R., Evans D., Beresford A., Bacon J., Friday A. (2010) Statistical Modelling and Analysis of Sparse Bus Probe Data in Urban Areas. In *13th International IEEE Conference on Intelligent Transportation Systems*, Madeira Island, Portugal, IEEE Intelligent Transportation Systems Society (September 2010), pp.1256-1263.

Carlo Berzuini, Statistical Laboratory

Causal inference in genetic epidemiology: looking into mechanism

We propose a method for the study of gene–gene, gene-environment and gene-treatment interactions which are interpretable in terms of mechanism. Tests for detecting mechanistic – as opposed to “statistical” – interactions have been previously proposed, but they are meaningful only if a number of assumptions and conditions are verified. Consequently, they are not always applicable and, in those situations where they are, their validity depends on an appropriate choice of stratifying variables. This paper proposes a novel formulation of the problem. We illustrate the method with the aid of studies where evidence from case-control studies of genetic association is combined with information from biological experiments, to elucidate the role of specific molecular mechanism (autophagy, ion channels) in susceptibility to specific diseases (Crohn’s Disease, Multiple Sclerosis).

Jack Bowden, MRC Biostatistics Unit

Optimal design and analysis procedures in two stage trials with a binary endpoint

Two-stage trial designs provide the flexibility to stop early for efficacy or futility, and are popular because they have a smaller sample size on average compared to a traditional trial with the same type I and II errors. This makes them financially attractive but also has the ethical benefit of reducing, in the long run, the number of patients who are given ineffective treatments. Therefore designs which minimise the expected sample size are referred to as ‘optimal’. However, two-stage designs can impart a substantial bias into the parameter estimate at the end of the trial. The properties of standard and bias adjusted maximum likelihood estimators, as well as mean and median unbiased estimators are reviewed with respect to a binary endpoint. Optimal two-stage design and analysis procedures are then identified that balance projected sample size considerations with estimator performance.

Chris J. Brien, Phenomics and Bioinformatics Research Centre, University of South Australia, Adelaide

Multiphase experiments in the biological sciences

Multitiered experiments are characterized as involving multiple randomizations (Brien et al., 2003; Brien and Bailey, 2006). Multiphase experiments are one class of such experiments, other classes being some superimposed experiments and some plant and animal experiments. Particularly common are multiphase experiments with at least one later laboratory phase (Brien et al., 2011); some examples of them, illustrating current research areas, will be presented.

References

- Brien, C.J. and Bailey, R.A. (2006) Multiple randomizations (with discussion). *Journal of the Royal Statistical Society, Series B (Methodology)*, 68, 571-609.
- Brien, C.J., Bailey, R.A., Correll, R.L., Harch, B.D., Payne, R.W. and Demtrio, C.G.B. (2003) Multitiered experiments web site. <http://chris.brien.name/multitier>
- Brien, C.J., Harch, B.D., Correll, R.L. and Bailey, R.A. (2011) Multiphase experiments with at least one later laboratory phase. I. Orthogonal designs. *Journal of Agricultural, Biological and Environmental Statistics*, available on-line at <http://dx.doi.org/10.1007/s13253-011-0060-z>.

Benilton Carvalho, Department of Oncology

On the exploration of Affymetrix ligation-based SNP assays

Single Nucleotide Polymorphisms (SNPs) are genetic variants that take place through the alteration of a single nucleotide (\mathcal{A} , \mathcal{C} , \mathcal{G} or \mathcal{T}) on the DNA sequence. At SNP locations, the possible nucleotides are referred to as alleles and are labelled, for simplicity, as A and B. The combination of alleles (AA, AB and BB) are called genotypes and play an important role on genome-wide association studies (GWAS), through which researchers investigate the association between traits of interest (like diseases) and genomic markers. GWASes depend strongly on the availability of accurate genotypes for the samples involved in the study. Several methodologies can be used for genotyping and one of the most common is DNA microarrays. Affymetrix is well-known for manufacturing one-color arrays comprised of 25 nucleotides long probes. However, their latest genotyping platform, called Axiom, uses a multicolor strategy to label the majority of the SNPs on ligation-based assays that use 30nt long probes.

Previous algorithms for both processing and genotype calling relied on the properties of the data generated by the older assays. Therefore, they may require modifications in order to be used with data from this new product. Here, we present a comprehensive investigation on the properties of the data generated using Axiom arrays, including the changes that are to be implemented on our algorithm for preprocessing SNP data and a discussion about the impact of this shift on downstream analyses involving SNP data.

Richard Durbin, Wellcome Trust Sanger Institute

Measures for capturing coverage of genetic variation in a population

The price of DNA sequencing and related technologies has dropped to a point where we can consider sequencing the genomes of a sufficiently large sample of individuals in a human population to capture almost all genetic variation.

We are already engaged in such studies in population isolates such as in Kuusamo in the north-east of Finland (population 20,000, founded by 34 families around 1650) or Orkney (population 15,000). There are many questions of efficient design, but also of what the quantity of interest is and how to measure it. Almost all variation is shared by inheritance, but every person has some genetic variants not inherited from our parents, due to new mutations. I will introduce some of the measures and strategies we are using, and I hope initiate discussion. I believe there is lots of scope for new ideas.

Robin Evans, Statistical Laboratory

Variation independent parametrizations

Variation independence can be a useful tool for developing algorithms and for parameter interpretation. We present a simple method for creating variation independent parametrizations of some discrete models using Fourier-Motzkin elimination, with some examples.

Bob Haining, Dept of Geography

Evaluating Peterborough's no cold calling initiative using space-time Bayesian hierarchical modelling

As part of a wider Neighbourhood Policing strategy, Cambridgeshire Constabulary instituted "No Cold Calling" (NCC) zones to reduce cold calling (unsolicited visits to sell products/services), which is often associated with rogue trading and distraction burglary. We evaluated the NCC-targeted areas chosen in 2005-6 and report whether they experienced a measurable impact on burglary rates in the period up to 2008. Time series data for burglary at the Census Output Area level is analysed using a Bayesian hierarchical modelling approach, addressing issues often encountered in small area quantitative policy evaluation. Results reveal a positive NCC impact on stabilising burglary rates in the targeted areas.

Ferenc Huszar and Neil MT Hounsby, Computational and Biological Learning Lab,
Dept of Engineering

Bayesian sequential experiment design for quantum tomography

Quantum tomography is a valuable tool in quantum information processing and experimental quantum physics, being essential for characterisation of quantum states, processes, and measurement equipment. Quantum state tomography (QST) aims to determine the unobservable quantum state of a system from outcomes of measurements performed on an ensemble of identically prepared systems. Measurements in quantum systems are non-deterministic, hence QST is a classical statistical estimation problem.

Full tomography of quantum states is inherently resource-intensive: even in moderately sized systems these experiments often take weeks. Sequential optimal experiment design aims at making these experiments shorter by adaptively reconfiguring the measurement in the light of partial data. In this talk, I am going to introduce the problem of quantum state tomography from a statistical estimation perspective, and describe a sequential Bayesian Experiment Design framework that we developed. I will report simulated experiments in which our framework achieves a ten-fold reduction in required experimentation time.

Chris Jackson, MRC Biostatistics Unit

Bayesian evidence synthesis to estimate progression of human papillomavirus

Human papillomavirus (HPV) types 16 and 18 are associated with about 70% of cervical cancers. To evaluate the long-term benefits of cervical screening and vaccination against HPV, estimates of the natural history of HPV are required. A Markov model has previously been developed to estimate progression rates of HPV, through grades of neoplasia, to cancer. The model was fitted to cross-sectional data by age group from the UK, including data from a trial of HPV testing, population cervical screening data, and cancer registry data. Parameter uncertainties and model choices were originally only acknowledged by informal scenario analysis. We therefore reimplement this model in a Bayesian framework to take full account of parameter and model uncertainty. Assumptions may then be weighted coherently according to how well they are supported by data. There is a complex network of evidence and parameters, involving misclassified and aggregated data, data available on different age groupings, and external data of indirect relevance. This is implemented as a Bayesian graphical model, and posterior distributions are estimated by MCMC. This work raises issues of uncertainty in complex evidence syntheses, and aims to encourage greater use in practice of techniques which are familiar in the statistical world.

David Knowles, Computational and Biological Learning Lab, Dept of Engineering

Inferring an individual's "physiological" age from multiple ageing-related phenotypes

What is ageing? One hypothesis is that ageing is global systemic degradation of multiple organ systems. Based on this assumption we propose a linear model which attempts to infer an individual's "physiological" age from multiple clinical measurements. Inference is performed using the variational Bayes approximation using the Infer.NET framework, a Microsoft Research project akin to WinBUGS. We apply the model to around 6000 individuals in the Twins UK study and look for gene expression levels and SNPs associated with ageing "delta": the difference between an individual's physiological and chronological age.

We propose an extension allowing non-linear variation of the clinical variables with age using a mixture of experts model. Finally we question whether a model with multiple dimensions of ageing might more closely resemble reality.

John Marioni, European Bioinformatics Institute

Statistical modeling of gene expression levels

Next-generation sequencing (NGS) technology has revolutionized our ability to assay the genome, transcriptome and epigenome of multiple different organisms. However, to ensure that the data generated can be utilized to answer pertinent biological questions, it is vital that appropriate statistical tools are developed.

In this talk, I will discuss some of the statistical challenges that arise when modeling gene expression measurements made using NGS. I will also discuss how current models will have to be extended and adapted as we move from studying gene expression levels measured across large populations of cells to measurements made at the single-cell level.

Sebastian Nowozin, Microsoft Research Cambridge

Statistical problems in computer vision

Computer vision is one of the many fields that successfully adopted machine learning for building predictive models. Yet, despite their success some of the fields' most popularly used models such as conditional random fields remain poorly understood theoretically and require approximations to be practical. I discuss a few of open theoretical and practical questions in these models in the computer vision context.

Roland Ramsahai, Statistical Laboratory

Identifying the effect of treatment on the treated

In the counterfactual literature, the effect of treatment on the treated (ETT) is often branded as the effect on the treated group. This definition of ETT is vague and potentially misleading because ETT is the effect on those who would normally be treated. A more transparent definition of ETT is given within the decision theoretic framework. The proposed definition of ETT is used to highlight misuse of terminology in the literature and discuss the types of studies that can be used for identifying ETT. Criteria for identifying ETT from observational data, when there are unobserved confounders, are given. The criteria are compared to those formulated within the counterfactual framework.

Richard Samworth, Statistical Laboratory

Log-concavity, nearest-neighbour classification, variable selection and the Statistics Clinic

I will give a brief overview of some of my current research interests. These include log-concave density estimation and its applications, optimal weighted nearest neighbour classification, and the use of subsampling to improve high-dimensional variable selection algorithms. Finally, I will advertise the Statistics Clinic, <http://www.statslab.cam.ac.uk/clinic> which meets fortnightly during term, and where anyone in the university can obtain free statistical advice.

Shaun R Seaman and Ian R White, MRC Biostatistics Unit

Inverse probability weighting with missing predictors of missingness or treatment assignment

Inverse probability weighting is commonly used in two situations. First, it is used in a propensity score approach to deal with confounding in non-randomised studies of effect of treatment on outcome. Here weights are inverse probabilities of assignment to active treatment. Second, it is used to correct bias arising when an analysis model is fitted to incomplete data by restricting to complete cases. Here weights are inverse probabilities of being a complete case.

Usually weights are estimated by regressing an indicator of whether the individual receives active treatment (in the first situation) or is a complete case (in the second) on a set of predictors.

Problems arise when these predictors can be missing. In this presentation, I shall discuss a method that involves multiply imputing these missing predictors.

So-Youn Shin, Ann-Kristin Petersen, Christian Gieger, Nicole Soranzo, Wellcome Trust Sanger Institute

Structural equation modeling analysis for causal inference from multiple omics datasets

Recent developments in technology allow us to collect multiple highly-dimensional 'omics' datasets from thousands of individuals in a highly standardized and unbiased manner. Open questions remain how best to integrate the multiple omics datasets to understand underlying biological mechanisms and infer causal pathways. We have begun exploring causal relationships between genetic variants, clinically-relevant quantitative phenotypes and metabolomics datasets using Structural Equation Modeling (SEM), applied to a subset of the common disease loci identified from genome-wide association studies.

We provide proof-of-principle evidence that SEM analysis is able to identify reproducible path models supporting association of SNPs to intermediate phenotypes through metabolomics intermediates. We address further challenges arising from the analysis of multiple omics datasets and suggest future directions including nonlinear model based approaches and the simultaneous dimension reduction (or variable selection) methods.

David Spiegelhalter, MRC Biostatistics Unit and Statistical Laboratory

Communicating and evaluating probabilities

I will talk about three related projects: (a) Getting children to express numerical confidence in their knowledge (b) Collaboration with the Met Office in an online weather game incorporating probabilistic forecasts. (c) Development of an online quiz.

Sinan Yildirim, Statistical Laboratory

Forward Smoothing and Online EM in changepoint systems

In this talk, I will focus on forward smoothing in changepoint systems, which are generally used to model the heterogeneity in the statistical data. After showing the SMC implementation of forward smoothing, I will show how we can perform the online EM algorithm for parameter estimation in changepoint systems.

Eleftheria Zeggini, Wellcome Trust Sanger Institute

Rare variant analysis in large-scale association and sequencing studies

Recent advances in whole-genome genotyping technologies, the availability of large, well-defined sample sets, and a better understanding of common human sequence variation, coupled with the development of appropriate quality control and analysis pipelines, have led to the identification of many novel common genetic determinants of complex traits. However, despite these successes, much of the genetic component of these traits remains unaccounted for. One largely unexplored paradigm which may contribute to this missing heritability is a model of multiple rare causal variants, each of modest effect and residing within the same functional unit, for example, a gene. Joint analysis of rare variants, searching for accumulations of minor alleles in individuals, for a dichotomous or quantitative trait, may thus provide signals of association with complex phenotypes that could not have been identified through traditional association analysis of single nucleotide polymorphisms (SNPs). However, statistical methods to perform such joint analyses of rare variants have not yet been fully developed or evaluated.

We have implemented rare variant analysis methods in user-friendly software and have extended approaches to collapsing rare allele tables and allele-matching tests by incorporating variant-specific quality scores (for example arising from next generation sequencing studies in which different positions have been covered at different depths) and genotype-specific probabilities (for example arising from 1000 genomes project-imputed data). We evaluated these methods and find increases in power to detect association under varying allelic architectures and parameters. We make recommendations for the analysis of rare variants in large-scale association and next generation sequencing studies.