

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

Titles and abstracts of posters

Elizabeth LC Merrall, Sheila M. Bird, Sharon J. Hutchinson, MRC Biostatistics Unit

High Drug-Related Death (DRD) rate soon after hospital-discharge for drug-treatment clients in Scotland, 1996-2006: record-linkage study

Record-linkage studies revealed that UK prisoners' drugs-related death rate was 7.5 times higher in the fortnight following release from prison than per-fortnight in the subsequent 10 weeks (95% CI: 5.7-9.9). We estimated that 1 in 200 adult prisoners with a history of heroin injection was dead from overdose within 2 weeks of release from Scottish prisons in the later 1990s. Naloxone is the heroin antidote and is administered by intramuscular injection. Since 2005, Naloxone can be administered by anyone in an emergency to save life. The Medical Research Council funded the pilot phase of a randomized controlled trial (N-ALIVE) which aims to randomize 56,000 prisoners with a history of heroin injection to receive, on release, an N-ALIVE pack which contains, or does not contain (no placebo) a syringe with Naloxone that, if located by a present-other, can be administered to them in the event of opiate-overdose. N-ALIVE Trial aim to test whether Naloxone-on-release delivers a 30% reduction in overdose deaths in the first 4 weeks after release and 20% reduction in the next 8 weeks.

Meanwhile, Scotland introduced, and funded, take-home Naloxone as a public health policy in 2011, and Wales has followed suit. On the basis of this poster's results, Scotland's chief medical officer has recommended that hospital-doctors consider the prescription of take-home Naloxone when discharging opiate-dependent clients. Trial, or policy? What do you think . . .

Simon Byrne and Philip Dawid, Statistical Laboratory

The structural Markov property

We consider the problem of model determination for undirected decomposable graphical models. Also referred to as structural learning, this involves inferring the structure of the underlying graph from the observed data. This problem has been the focus of much recent work in both Bayesian and frequentist statistics, particularly in the context of covariance matrix estimation and contingency table analysis.

A Bayesian approach to this problem requires the specification of a probability distribution over a set of graphs. We introduce a “structural Markov property” for such distributions: the structure of induced subgraphs should be independent conditional on the existence of a complete separating subgraph.

The form of the structural Markov property is analogous to the Markov property of the sampling distribution, and the hyper Markov properties of the parameter distributions. By exploiting these properties we can obtain an efficient MCMC methods for determining the posterior based on local computations.

Furthermore, the structural Markov property characterises an exponential family over the set of decomposable graphs, which forms the conjugate prior for sampling from a family of compatible Markov distributions.

Stephen Burgess, School of Mathematical Sciences, Queen Mary, University of London

Mendelian randomization: the use of genetic variants as an instrumental variable for assessing causal associations in observational data

Mendelian randomization is an epidemiological method which uses genetic variation to estimate the causal effect of the change in a modifiable risk factor on an outcome from observational data. A genetic variant satisfying the assumptions of an instrumental variable for the risk factor of interest divides a population into subgroups which differ systematically only in the risk factor. A causal estimate can be calculated which is asymptotically free of bias from confounding and reverse causation. This poster provides a general introduction to instrumental variables and Mendelian randomization, illustrating with the example of the effect of “bad” cholesterol (LDL-C) on coronary heart disease (CHD) levels. The illustration is extended to motivate potential future work, integrating data on multiple genetic variants affecting a range of risk factors and simultaneously estimating several causal effects in a single dataset.

Silvia Chiappa, Microsoft Research Cambridge

Computationally efficient extension of fastPHASE for genotype and haplotype estimation in trios

Incorporating family structure, such as trio (mother-father-child) structure, in methods for imputing missing genotypes and reconstructing haplotypes can lead to substantial performance improvement with respect to considering individuals as unrelated. However, this comes with increased computational burden and approximations are normally required. In this poster, we present an extension of fastPHASE that accounts for trio structure. Computational complexity is reduced to the same order as in the basic fast-PHASE model by employing a variational approximation.

Panayiota Constantinou and Philip Dawid, Statistical Laboratory

Conditional Independence in the Decision-Theoretic Framework and identification of causal quantities

The calculus of conditional independence arises to be the basic element which enables us to express causal concepts in the Decision-Theoretic framework of Statistical Causality. In this framework, we differentiate between observational and interventional regimes using a non-stochastic variable (to index the regimes) and use the notion of conditional independence to state conditions under which we can relate them. Here, we rigorously extend the language and calculus of conditional independence to incorporate stochastic and non-stochastic variables and use the properties that accrue to identify causal quantities. More specifically, we present under which conditions we can identify the Consequences of Dynamic Treatment Strategies.

Philip Dawid¹ and **Monica Musio**², 1Statistical Laboratory, 2Dipartimento di Matematica, Università degli Studi di Cagliari

Local scoring rules for spatial processes

We display Besag's pseudo-likelihood as a special case of a general estimation technique based on proper scoring rules. Such a rule supplies an unbiased estimating equation for any statistical model, and this can be extended to allow for missing data. When the scoring rule has a simple local structure, as in many spatial models, the need to compute problematic normalising constants is avoided. We illustrate the approach through an analysis of data on disease in bell pepper plants.

Audrey Q. Fu, Department of Physiology, Development and Neuroscience

Inferring the transcriptional mechanism from genomic data

Mark Haggard¹, Helen Spencer², Jan Zirk-Sadowski¹, 1Department of Experimental Psychology, 2Multi-centre Otitis Media Study Group

Enhancing information value and causal inference in structural equation modelling (SEM) via a matrix of designed model contrasts

Many applications of SEM in biology, psychology and social science are only conceptual summaries: descriptions of covariance between complex sets of variables. These use neither hypothesis tests on paths of interest (ie in the controlled SEM context) nor clear contrasts documenting the information value of the model relative to less interesting alternative models with the same number of serial stages. Parsimony-adjusted fit indices are a move in the right direction, but do not quite exhaust the epistemological concept of explanatory power, which is helped by the idea of a "worthy opponent" model for the preferred model. We have modelled a dataset on children's ear problems, development and quality of life to make clear the major channels of mediation. The preferred multi-channel parallel model fits the data well, but more importantly, it performs better than a less interesting serial-only model with the same number of serial stages (worthy opponent). So there are two main processes of influence, not one. We also show that a very similar path structure emerges when the covariance of early stages is driven not by spontaneous individual differences at baseline, but by randomised surgical treatment of the ear problems (ie an experimental design, giving a hybrid paradigm, as regards the mode of control). This enhances validity and generalisability of the whole class of model. Overall, such quasi-experimental modelling strategies seem to offer much to disciplines where the main analytical tools are based on covariance analysis.

Hilal Kazan, Microsoft Research Cambridge

Detailed binding preferences of RNA-binding proteins inferred from large-scale binding assays

RNA binding proteins (RBPs) play critical roles in post-transcriptional regulation but their target binding preferences remain largely uncharacterized. New large-scale binding assays (e.g. PAR-CLIP, RIP-seq, RNAcompete) support the construction of detailed models of RBP binding preferences and promise to rapidly expand our knowledge of RBP binding targets. However, these new binding data present a challenge to existing motif finding methods, as few methods are able to handle the size, detail or semi-quantitative nature of these data. To address this absence, we introduce MaLaRKey (Multilinear regression RNAcontext) a new motif finding method that uses a feature-based product model to represent RBP binding affinity for a given site. MaLaRKey motif models are fit to RNA binding data from large-scale assays by using multilinear regression to maximize the agreement between the RNA sequence affinity predicted by the motif model and that measured by the binding assay. MaLaRKey is able to recover sequence and structure binding preferences for numerous RBPs (e.g. Vts1, Khd1).

Radoslaw P. Lach, Mathew Garnett, Ultan McDermott, David J. Adams, Wellcome Trust Sanger Institute

Inferring anti-cancer drug resistance factors

Cancer is a genetic disease characterized by a loss of growth control. Although our knowledge of cancer has grown enormously during the last two decades treatment options are still limited, and for many cancer types the prognosis remains dismal. New sequencing technologies now make it possible to thoroughly characterize the genome of a cancers and potentially identify gene mutations that drive tumorigenesis. I analyzed over 300 cancer cell lines responses to set of commonly used cancer therapeutics. Here I show how a deep characterization of cancer genome and transcriptome can be applied to deduce factors influencing drug resistance

Y. Perez-Riverol^{1,2}, A. Sánchez, Y. Ramos¹, A. Schmidt³, M. Müller³, L. Betancourt¹, L. J. Gonzalez¹, R. Vera¹, G. Padron¹, V. Besada¹, 1Department of Proteomics, Center for Genetic Engineering and Biotechnology, Ciudad de la Habana, Cuba, 2EMBL Outstation, European Bioinformatics Institute, 3Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

In silico analysis of accurate proteomics, complemented by selective isolation of peptides

Protein identification by mass spectrometry is mainly based on MS/MS spectra and the accuracy of molecular mass determination. However, the high complexity and dynamic ranges for any species of proteomic samples, surpass the separation capacity and detection power of the most advanced multidimensional liquid chromatographs and mass spectrometers. Only a tiny portion of signals is selected for MS/MS experiments and a still considerable number of them do not provide reliable peptide identification. In this article, an in silico analysis for a novel methodology of peptides and proteins identification is described. The approach is based on mass accuracy, isoelectric point (pI), retention time (t(R)) and N-terminal amino acid determination as protein identification criteria regardless of high quality MS/MS spectra. When the methodology was combined with the selective isolation methods, the number of unique peptides and identified proteins increases. Finally, to demonstrate the feasibility of the methodology, an OFFGEL-LC-MS/MS experiment was also implemented. We compared the more reliable peptide identified with MS/MS information, and peptide identified with three experimental features (pI, t(R), molecular mass). Also, two theoretical assumptions from MS/MS identification (selective isolation of peptides and N-terminal amino acid) were analyzed. Our results show that using the information provided by these features and selective isolation methods we could found the 93MS/MS with false-positive rate lower than 5

Continuous logarithmic plots

Logarithmic plots are a common way to visualize data with a high dynamic range spanning multiple orders of magnitude. Despite the obvious advantages and its popularity, logarithmic plots show several defects:

- Zeroes cannot be handled
- Negative values cannot be handled
- Values close to zero are stretched out unproportionally (which might or might not be desired)

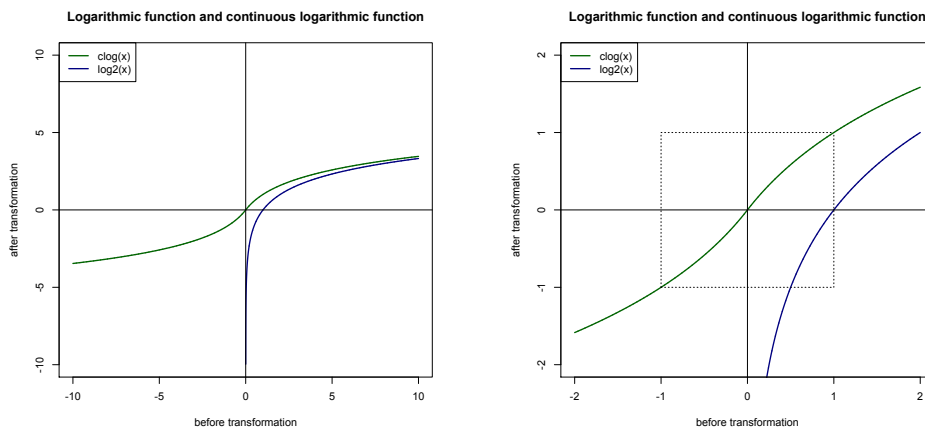
To address these problems, we present an alternative visualization method named continuous logarithmic plot (or fisheye plot), which exhibits the following properties:

1. It is a continuous function from ∞ to ∞
2. It is anti-symmetric, transforming negative values into negative values, positive values into positive ones and zero to zero
3. It is asymptotically logarithmic

Thus the plot is effectively logarithmic for large numbers, but handles small and negative numbers more gracefully. This property also allows linear shifting and scaling of the transformation effectively "zooming in" to high-density areas of the data and "blending out" the long tails of the distribution, acting in effect as a virtual fisheye lens. The transformation can be applied to either the X or the Y axis, or both axes at the same time. The basic formula for the continuous logarithm is the following:

$$clog(x) = sgn(x) * log_2(abs(x) + 1)$$

Using the logarithm with base two ensures that the values -1, 0, and 1 all transform to themselves. The following plot shows the function against the traditional (base 2) logarithmic function in various ranges:



Martin Szummer, Microsoft Research Cambridge

Semi-supervised learning to rank with preference regularization

We propose a semi-supervised learning to rank algorithm. It learns from both labeled data (pairwise preferences or absolute labels) and unlabeled data. The data can consist of multiple groups of items (such as queries), some of which may contain only unlabeled items. We introduce a preference regularizer favoring that similar items are similar in preference to each other. The regularizer captures manifold structure in the data, and we also propose a rank-sensitive version designed for top-heavy retrieval metrics including NDCG and mean average precision.

The regularizer is employed in SSLambdaRank, a semi-supervised version of LambdaRank. This algorithm directly optimizes popular retrieval metrics and improves retrieval accuracy over LambdaRank, a state-of-the-art ranker that was used as part of the winner of the Yahoo! Learning to Rank challenge 2010. The algorithm runs in linear time in the number of queries, and can work with huge datasets.