DSTS Two-day meeting

Department of Mathematical Sciences
University of Copenhagen
19–20 May 2015

Theis Lange, Department of Biostatistics, University of Copenhagen

Disentangling causal pathways using natural effects models – better and faster mediation analysis

Important questions within the fields of social science, epidemiology as well as clinical trial research involve decomposing the effect of an interventions or exposure into direct and indirect effects working through a defined mediator; hopefully leading to a better understanding of the underlying mechanisms. In this talk I present the class of natural effects models, which allows causal inference based mediation analysis to be understood (and performed) in terms of traditional regression analyses. I will introduce the model class and the interpretation of the parameters in the model. Next I will present a just published R-package, which makes estimation of natural effects models both simple and intuitive. Finally, I explain how natural effects models can be used to address more complex mediation questions; in particular multiple mediators.

Eva Löcherbach, Département de Mathématiques, Université de Cergy–Pontoise, France

First steps towards statistical analysis in stochastic Hodgkin–Huxley models

This is a joint work with R. Höpfner, Mainz, and Michèle Thieullen, Paris.

The deterministic Hodgkin–Huxley model for the membrane potential of a single neuron describes the mechanism of spike generation (spikes=emission of action potentials) in response to an external input. We study a stochastic version of this model in which a cortical neuron receives some T–periodic (unknown) signal S from its dendritic system. In this frame, the stochastic Hodgkin–Huxley model is a coupled system of diffusion equations describing the observed membrane potential process (first coordinate) as well as unobserved coordinates which model ion currents. Observing the first coordinate alone leads to a non–Markovian process.

The main interest of modern neurosciences is to understand how neurons respond to external stimuli. Therefore, it is important to build statistical procedures aiming at estimating the unknown signal S (or some important features of S, for instance the unknown periodicity T), based on the observation of the membrane potential process.

In our work we establish "periodic ergodicity" of the process, based on a detailed study of the transition densities of the stochastic Hodgkin–Huxley model. The main difficulty comes from the fact that the model is a highly degenerate diffusion with time inhomogeneous coefficients. Moreover we obtain limit theorems for the sequence of successive spike intervals which allow to describe the spiking activity in the long run.

Philip Hougaard, Lundbeck

Interval censoring in multivariate survival data and multi-state survival data

Interval censoring means that an event time is only known to lie in an interval $(L, R]$, with L the last examination time before the event, and R the first after. In the univariate case, parametric models are easily fitted, and for non–parametric models, the mass is placed on some intervals, derived from the L and R points. Asymptotic results are simple for the former and complicated for the latter. This paper reviews extensions in two directions. One case is multivariate survival data, like eruption times for teeth examined at visits to a dentist or times of events for related individuals. Parametric models extend easily to multivariate data.

However, non–parametric models are intrinsically more complicated. It is difficult to find the intervals with positive mass and estimated interval probabilities may not be unique. A semi-parametric model makes a compromise, with a parametric model, like a frailty model, for the dependence and a non-parametric model for the marginal distribution. These three models are compared and discussed. The semi–parametric approach may be sensible in many cases, as it is more flexible than the parametric models, and it avoids some technical difficulties with the non–parametric approach. The other case is multi–state models, in practice exemplified by two cases, the illness-death model (where the time of illness occurrence is interval censored and the death time is known precisely) and the case known as double censoring (referring to analysis of the time between two events, which are both subject to censoring, like an incubation time, where the time of becoming infected is interval censored). Also in this case, it may be preferable to use models that are more restrictive than those we would have used in case the data had only been subject to right censoring.

Philip Dawid, Statistical Laboratory, University of Cambridge, UK

Theory and Applications of Proper Scoring Rules

Suppose you have to quote a probability distribution Q for an unknown quantity X. When the value x of X is eventually observed, you will be penalised by an amount $S(x, Q)$. The function S is a *proper scoring rule* if, when you believe X has distribution P, your expected penalty $S(P, Q) := \mathrm{E}_{X\sim P} S(X, Q)$ is minimised

by the honest quote $Q = P$.

Proper scoring rules can be used, as above, to motivate you to assess your true uncertainty honestly, as well as to measure the quality of your past probability forecasts in the light of the actual outcomes. They also have many other statistical applications. I will discuss some characterisations, properties and specialisations of proper scoring rules, and describe some of their uses, including robust estimation and Bayesian model selection.

Martin Bøgsted, Department of Clinical Medicine, Aalborg University and Department of Haematology, Aalborg University Hospital

*Integrating Multiple High Dimensional Data Sets with an Application in Lymphoma Cancer Treatment Research**

For a long time costly low sample high–dimensional data sets have fuelled research in statistical techniques handling data with much more features than samples. However, the emerging development of vast online repositories of publicly available data now increases sample size with an unprecedented speed. Take as an example the Gene Expression Omnibus (GEO), which is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data sets. Since its introduction in 2000 the database has increased to 55,429 datasets containing more than 1,350,000 million samples. This calls for the development of robust and flexible tools to integrate results of multiple high–dimensional experiments across laboratory batch effects and varying biotechnological technologies. By use of publicly available data sets from GEO we will illustrate how we have developed methods for integrating data across laboratories and technologies to validate and study new sub–classifications of lymph node cancer. The talk will cover the whole process from the preclinical models to the individual assignment and classification of tumors to be included into early clinical phase I–II trials.

*The research is funded by the National Experimental Therapy Partnership (NEXT), which is financed by a grant from Innovation Fund Denmark.

Therese Graversen, Department of Mathematical Sciences, University of Copenhagen

Statistical interpretation of forensic DNA mixtures

DNA evidence is used in forensic casework for the identification of individuals. When a sample contains DNA from multiple individuals, the interpretation of a sample is more challenging because the individual DNA profiles are observed only through a mixed signal in the associated electropherogram (EPG). Interpretation

of a mixed sample relies in part on statistical methods, and these are typically developed to target a specific question of interest.

However, a more unified framework may be based on a joint statistical model for the observed EPG and the underlying DNA profiles. This allows a wide range of questions to be addressed entirely within a single framework, avoiding the repeated re–invention of methodology and improving both the consistency and the clarity of reasoning about the sample.

Importantly, the model–based approach enables the scientist to perform systematic model checking, and I will give some examples to illustrate the value of model checking both to a development process and to casework.

Concurrent development of computational methodology has ensured that the approach is also practically feasible; a full implementation has been made available to the forensic community through the R package DNAmixtures.

Grégory Nuel, INSMI, Stochastics and Biology Group (PSB), LPMA, UPMC, Sorbonne University, Paris France

Fast estimation of posterior probabilities in change-point models through a constrained hidden Markov model

The detection of change–points in heterogeneous sequences is a statistical challenge with applications across a wide variety of fields. In bioinformatics, a vast amount of methodology exists to identify an ideal set of change–points for detecting Copy Number Variation (CNV). While considerable efficient algorithms are currently available for finding the best segmentation of the data in CNV, relatively few approaches consider the important problem of assessing the uncertainty of the change–point location. Asymptotic and stochastic approaches exist but often require additional model assumptions to speed up the computations, while exact methods have quadratic complexity which usually are intractable for large datasets of tens of thousands points or more. In this paper, we suggest an exact method for obtaining the posterior distribution of change-points with linear complexity, based on a constrained hidden Markov model. The methods are implemented in the R package postCP, which uses the results of a given change-point detection algorithm to estimate the probability that each observation is a change–point. We present the results of the package on a publicly available CNV data set ($n = 120$). Due to its frequentist framework, postCP obtains less conservative confidence intervals than previously published Bayesian methods, but with linear complexity instead of quadratic. Simulations showed that postCP provided comparable loss to a Bayesian MCMC method when estimating posterior means, specifically when assessing larger–scale changes, while being more computationally efficient. On another high–resolution CNV data set ($n = 14, 241$), the implementation processed information in less than one second on a mid-range laptop computer.