Ad-hoc BBS Seminar: Missing Data and Graphical Models

Date: Monday Oct 17 2016, Time: 15.00-17.00
Venue: Roche IT Training Center / Aeschentor, Aeschenvorstadt 56, 4051 Basel

The BBS is proud to invite to an ad-hoc seminar with focus on Missing data and Graphical Models with Speakers from University of Warwick, Novartis and Roche.

The seminar is free of charge.

Program:

15.00-15.40
Jane Hutton, Uni Warwick — ***"Missing data and how to see biased results using Chain Event Graphs"***

15.40-16.20
Giusi Moffa, Novartis — ***"Cancer profiling and subtype discovery with Bayesian inference for acyclic digraphs"***

16.20-17.00
Markus Elze, Roche — "***Propensity scores methods and covariate adjustment in practice"*** "

We look forward to your participation.

Kind Regards,
Dominik Heinzmann, on behalf of the BBS
(questions, you reach me at ***dominik.heinzmann@roche.com*****)**

Abstracts

Jane Hutton, University of Warwick
"Missing data and how to see biased results using Chain Event Graphs"

Any study of interventions which take time to have an effect, whether treatment for cancer or diets for weight loss, require measurements, often several measurements to be recorded at some interval after the initial assessment. For some conditions, such as diabetes, it is reasonable to expect people to visit a clinic regularly, but for minor conditions it is usual to rely on mail or telephone questionnaires. Study participants might be too busy to complete the questionnaire, might think their reply is not important, might have moved,

or might have died. What are the possible biases in our conclusions if we summarise only the available data? If the reasons for non-response are not related to the study, our results might be unbiased. If men recover more quickly than women, and independent of recovery are less likely to respond, simple means would under-estimate the time to recovery. We could adjust the results for the known probability of responding, and with a randomised study, we can find a good estimate of the difference between treatments.

However, if one treatment leads to quicker recovery, and those who have recovered are less likely to respond, then it is very difficult to obtain a reliable estimate of treatment effect.

This talk will begin with common definitions of missing data, which reflect the situations described above. Chain Event Graphs will then be introduced: they are statistical models for a set of random variables whose joint probability function is described in terms of a graph. The graphs are derived from probability trees by merging nodes in tree with the same associated conditional probabilities. This results in an accessible, visual representation of the statistical model. The concepts will be illustrated using data from a randomized controlled trial of immobilization after acute sprain. The trial recruited 584 patients who were randomized to tubular bandage,`Bledsoe boot', ankle brace or 10-day below knee cast. Ankle function was measured using a questionnaire prior to randomization and at four, twelve and 39 weeks after randomization. The response rates were 83%, 82% and 76% respectively. Is the data which is missing important?

Giusi Moffa, Novartis

"Cancer profiling and subtype discovery with Bayesian inference for acyclic digraphs"

The progression of cancer may be viewed as an evolutionary process in which the accumulation of genomic alterations leads to certain capabilities of the cancer tissue that are essential for example to evade the host immune system and continue growth. Which genomic alterations are beneficial for tumour progression at a given point in time depends, however, on previous alterations and the environment of the cancer cell. Understanding these dependencies will improve cancer driver identification, elucidate mechanisms of resistance, and open new opportunities for personalised cancer treatment.

A number of models have been proposed to estimate these dependencies from cross-sectional data. Most of these (e.g. [1,2]) limit the structure of dependencies to trees and estimate the probability of an event given its single parent event. Conjunctive Bayesian Networks (CBNs) [3] consider a subset of directed acyclic graphs (DAGs) to model dependencies. The underlying assumption of CBNs is, that a mutation can only occur once all the mutations it depends on (parent events) have occurred.

We relax this strict assumption and for each gene we fit parameters for each state of its parent events over the complete space of DAGs. Other than for tree structures, inference of Bayesian networks is in general computationally demanding. For DAGs we utilise our novel algorithm based on their combinatorial structure [4] and our recent computational extension to large networks. This allows inference for much larger gene networks than for previous methods for CBNs limited to about 15 nodes.
Therefore, more plausible biological models can be used to estimate likely progression paths for genetic data from large cohorts like the TCGA [5]. We estimate progression models for the 12 main TCGA cancer types, highlighting the main genes involved, and analyse the similarities and differences of cancer progression across samples. Furthermore, we can cluster cancer patient samples based on similarities derived from our probabilistic graphical models.

[1] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics, 21:2106--2107, 2005.
[2] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of Computational Biology,6:37--51, 1999.
[3] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel. Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics, 25:2809--2815, 2009.

[4] J. Kuipers and G. Moffa. Partition MCMC for inference on acyclic digraphs. Journal of the American Statistical Association. DOI:10.1080/01621459.2015.1133426
[5] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Mills Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45:1113--1120, 2013.

Markus Elze, Roche

"Propensity scores methods and covariate adjustment in practice"

Propensity scores (PS, Rosenbaum & Rubin, 1983) are increasingly used to estimate causal effects in observational studies. The PS is the probability of an experimental unit being assigned to a "treatment" given a set of covariates. PS allow the estimation of the treatment effect without dealing with potential confounding. When the number of observations in one "treatment" group is small, this approach may perform better than covariate adjustment. However, there is also a potential for overfitting in PS models (Senn, Graf, Caputo, 2007).

Popular PS methods include matching or stratifying observations based on the PS as well as inverse probability weighting (IPW). An important development was the advent of "doubly robust" methods (Robins, Rotnitzky, Zhao, 1994) that offer some robustness to model misspecification either in the PS or the final treatment model. In the first part of this talk, we will discuss the concepts behind these methods and offer a brief introduction to PS methods.

Then we will compare results of covariate adjustment and PS methods for time-to-event data from four different cardiovascular datasets: the ADAPT-DES prospective, multicentre registry (Stone, Witzenbichler, Weisz, et al. 2013), CHARM programme (Pocock, Wan, Pfeffer, et al. 2006), PROMETHEUS cohort study (Mehran, Baber, Sartori, et al. publication forthcoming), and the THIN population-based cohort study (Smeeth, Douglas, Hall, et al. 2009). We will demonstrate advantages and limitations of the methods and highlight where they produce different results. In particular, we will discuss the role of extreme PS, trimming, and robust estimation.