UNIVERSITY OF COPENHAGEN

Section of Biostatistics

NBBC17 program

Time	Monday June 19th		Tuesday June 20th		Wednesday June 21th	
8.00-9.00	Registration		Registration			
9.00-9.30	Welcome Elizabeth Thompson IBS President					
9.30-10.30	Keynote 1 (35.01.06)		SJS Lecture Keynote 2 (35.01.06)		Keynote 3 (35.01.06)	
10.30-11.00	Coffee break		Coffee break		Coffee break	
11.00-12.30	I3 (35.01.06)	C4 (35.3.12)	I5 (35.01.06)	C2 (35.3.12)	I1 (35.01.06)	C3 (35.3.12)
12.30-14.00	Lunch		Lunch		Lunch	
14.00-15.30	I4 (35.01.06)	C1 (35.3.12)	I2 (35.01.06)	C5 (35.3.12)		
15.30-16.00	Coffee break		Coffee break			
16.00-17.30	C6 (35.01.06)	C7 (35.3.12)	C8 (35.01.06)	C9 (35.3.12)		
17.30-	Welcome reception		18.30- Conference dinner			

Invited Session 1 Digital epidemiology (June 21th, room 35.01.06, chair: Jaakko Nevalainen)

11.00-11.30 Juha Karvanen: External information helps to reduce selection bias in health examination surveys

11.30-12.00 Krista Lagus: To be announced

12.00-12.30 Ingemar J. Cox: Inferring health information from non-health sources

Invited Session 2 Applied probability models in biosciences (June 20th, room 35.01.06, chair: Kristi Kuljus)

14.00-14.30 Brona Brejová: Inference criteria for hidden Markov models in biological sequence analysis

14.30-15.00 Stephané Robin: Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree

15.00-15.30 Mae Woods: Mathematical modelling and Bayesian inference reveals new insights into structural variation of the human genome

Invited Session 3 Functional data analysis (June 19th, room 35.01.06, chair: Helle Sørensen)

11.00-11.30 R. Todd Ogden: Functional Data modeling of dynamic PET data

11.30-12.00 Bo Markussen: Markov Component Analysis and Functional Regression

12.00-12.30 Simone Vantini: Advances in local inference for functional data with application to tongue profile analysis

Invited Session 4 Statistics in climate science (June 19th, room 35.01.06, chair: Sara Sjöstedt de Luna)

14.00-14.30 Gudrun Brattström: Climate time series: signal, noise and additivity

14.30-15.00 Johan Lindström: Reconstructing past landscapes - space-time modelling of compositional data with zero counts

15.00-15.30 James Sweeney: Adressing the statistical challenges of estimating past climate from pollen sources

Invited Session 5 Longitudinal and time to event data (June 20th, room 35.01.06, chair: Jon Michael Gran)

11.00-11.30 Cécile Proust-Lima: Joint modelling of multiple latent processes and a clinical endpoint: application in Alzheimer's disease

11.30-12.00 Kjetil Røysland: Causal validity of marginal structural models for event history data and causal local independence graphs

12.00-12.30 Niels Keiding: Survival analysis around a cross-section and unobserved heterogeneity

---

Contributed session 1 Causal inference (June 19th, room 35.3.12, chair: Theis Lange)

14.00-14.30 Michal Abrahamowicz: Controlling for Unmeasured Confounding in Time-to-Event Analyses of Treatment Effects: the Missing Cause Approach

14.30-15.00 Per Kragh Andersen: Causal Inference in Survival Analysis using Pseudo-observations

15.00-15.30 Pål Christie Ryalen: Causal inference in survival analysis – continuous weighs and an application to cancer registries

Contributed session 2 Modern regression modeling (June 20th, room 35.3.12, chair: Magne Thoresen)

11.00-11.22 Ann-Sophie Buchardt: Identifying Interactions via Hierarchical Lasso Regularization

11.23-11.45 Riccardo De Bin: Strategies to handle Mandatory Covariates using Model- and Likelihood-based Boosting

11.45-12.07 Razaw Al-Sarraj: Generalized prediction of random effects in balanced and unbalanced two-factorial random-effects models

12.08-12.30 Gilles Guillot: Predicting Dermal Absorption of Chemicals from Physico–chemical Properties and Experimental Conditions. A Generalized Linear Mixed Model Approach

Contributed session 3 Applications (June 21th, room 35.3.12, chair: Sven Ove Samuelsen)

11.00-11.30 Aldana Rosso: Risk factors for treatment discontinuation for neovascular AMD

11.30-12.00 Michael Höhle: Evaluation of Quality of Care using Patient Follow-up Data

12.00-12.30 Rune Hoff: The impact of high-school completion - a multi-state model for work, education and health in young men

Contributed session 4 Survival analysis (June 19th, room 35.3.12, chair: Per Kragh Andersen)

11.00-11.30 Maja Pohar Perme: A Pseudo-observations Estimator in Relative Survival Analysis

11.30-12.00 Christian Pipper: Semiparametric multi-parameter regression survival modelling

12.00-12.30 Essi Syrjälä: Joint modelling on early nutrition and advanced beta-cell autoimmunity: A comparison among different approaches

Contributed session 5 Generalized linear mixed models and R (June 20th, room 35.3.12, chair: Helle Sørensen)

14.00-14.30 Johannes Forkman: Assessing precision in coefficients of variation between and within subjects using generalized pivotal quantities

14.30-15.00 Lars Rönnegård: Data Analysis using Hierarchical Generalized Linear Models with R

15.00-15.30 Hans J. Skaug: TMB: a flexible framework for developing mixed model software in R

Contributed session 6 Models for/with hidden structures (June 19th, room 35.01.06, chair: Jacob Hjelmborg)

16.00-16.30 Kristi Kuljus: Comparison of hidden Markov chain models and hidden Markov random field models in estimation of computed tomography images

16.30-17.00 Sara Sjöstedt de Luna: Multi-resolution Clustering of Time Dependent Functional Data with Applications to Climate reconstruction

17.00-17.30 Morten Valberg: The surprising implications of familial association in disease risk

Contributed session 7 Genomic studies (June 19th, room 35.3.12, chair: Claus Ekstrøm)

16.00-16.22 Krista Fischer: Some statistical challenges on the road from GWAS to personalized risk prediction

16.23-16.45 Valeria Vitelli: A rank-based Bayesian approach to combining genomic studies

16.45-17.07 Sarunas Germanas: A Novel Variant Recalibration Method for Detection of Rare Mutations in Next-Generation Sequencing Experiments

17.08-17.30 Elizabeth Thompson: Estimation of Pairwise Relatedness and Joint IBD

Contributed session 8 Time series analysis and prediction (June 20th, room 35.01.06, chair: Theis Lange)

16.00-16.30 Aksel Karl Georg Jensen: Estimating Natural Direct Effects in RCT'S With Multiple Time-varying Intermediates

16.30-17.00 Xingwu Zhou: Intervention time series analysis with its application to the Swedish tobacco quitline

17.00-17.30 Xiaoran Lai: Personalized Computer Simulations of Breast Cance Tumors treated with Neoadjuvant Chemotherapy with and without Bevacizumab: A Proof-of-concept

Contributed session 9 Statistical methodology (June 20th, room 35.3.12, chair: Esben Budtz-Jørgensen)

16.00-16.30 Tetiana Gorbach: Estimating partial correlation with data missing not at random

16.30-17.00 Jacob B. Hjelmborg On the Linfoot Informational Correlation

17.00-17.30 Jan-Eric Englund: Methods for quantifying fruit dehiscence

OBJECT ORIENTED DATA ANALYSIS

J. S. Marron

Department of Statistics and Operations Research, University of North Carolina

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. In medical image analysis, the data objects are often shapes, which naturally lie on manifolds. A series of successive improvements in the statistical analysis of shape data objects, developed through deep combination of statistical ideas with differential geometry, together with their practical implications is described. If time permits, an application of topological data analysis to a set of tree structured data objects will also be considered.

Keywords: Manifold Data, Object Oriented Data Analysis, Principal Component Analysis, Shape Statistics

NBBC17 **Keynote 2: SJS Lecture (June 20th, room 35.01.06, chair: Thomas Scheike)**

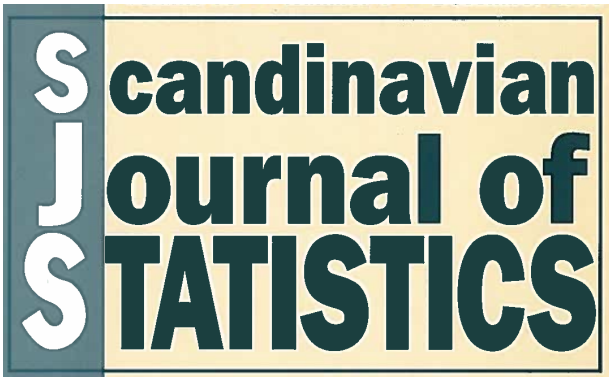
LEARNING FROM A LOT: EMPIRICAL BAYES IN HIGH-DIMENSIONAL PREDICTION SETTINGS

Wiel, Mark van de

Vrije Universiteit, Amsterdam

Empirical Bayes is a versatile approach to 'learn from a lot' in two ways: first, from a large number of variables and second, from a potentially large amount of prior information, e.g. stored in public repositories. We review applications of a variety of empirical Bayes methods to a broad spectrum of prediction methods including penalized regression, random forest, linear discriminant analysis, and Bayesian models with sparse or dense priors. We discuss 'formal' empirical Bayes methods which maximize the marginal likelihood, but also more informal approaches based on other data summaries. Empirical Bayes is contrasted to cross-validation and full Bayes. Hybrid approaches will be discussed.

Empirical Bayes is particularly useful when the prior or penalty contains multiple parameters that model a priori information on the variables, termed 'co-data'. In practice, one often seeks for predictors that require measuring a few variables only. We will show that in such a context the combination of empirical Bayes and co-data can render a large improvement in predictive performance and typically stabilizes variable selection. Methods will be illustrated on several cancer genomics applications. Finally, we shortly discuss extensions to other problems such as network reconstruction and drug target models.



NBBC17 **Keynote 3 (June 21th, room 35.01.06, chair: Per Kragh Andersen)**

FAMILIAL AGGREGATION OF AGE-OF-ONSET FOR SPECIFIC DISEASES

Thomas Scheike

University of Copenhagen, Denmark

In this talk I will cover several approaches for trying to estimate familial aggregation for time to event data, that could be time to death or more typically competing risks. One can model these data on either the hazards scale and on the the cumulative incidence scale. Both these modeling approaches will recover the same information but of course are quite different. Random effects models in these settings are very useful but difficult to use in practice because of limited software, in particular for the often large data that could be available via for example registries.

Given that familial aggregation has been established there is often a desire to say something about a possible time-structure in the dependence or if the timing could be related to the risk level. For some cancers it is believed that early cancers are high risk. I will show how some of these questions can be addressed.

EXTERNAL INFORMATION HELPS TO REDUCE SELECTION BIAS IN HEALTH EXAMINATION SURVEYS

Juha Karvanen

University of Jyvaskyla

The decreasing participation rates and selective non-participation peril the representativeness of health examination surveys (HES). In this talk I will review the results from the project „Non-participation in health examination surveys" (NoPaHES, `http://www.ehes.info/nopahes/`) and provide some recommendations for the design, implementation and analysis of HES. The distinctive feature of the NoPaHES project is the coverage of the register data. The register linking is carried out for both participants and non-participants to obtain data on deaths, hospitalizations, socio-economic status and medical reimbursements. These register data can be used to analyze the characteristics of non-participants, and to reduce selection bias. The use of follow-up data in the estimation of the prevalence of daily smoking and heavy alcohol usage is considered as an example.

References:

Karvanen J., Tolonen H., Härkänen T., Jousilahti P., Kuulasmaa K. (2016) Selection bias was reduced by recontacting non-participants, Journal of Clinical Epidemiology, 76, 209-217.

Härkänen T. , Karvanen J. , Tolonen H. , Lehtonen R., Djerf K., Juntunen T., Koskinen S. (2016) Systematic handling of missing data in complex study designs - experiences from the Health 2000 and 2011 Surveys. Journal of Applied Statistics, 43:15, 2772-2790.

Kopra J., Karvanen J., Härkänen T. (2016) Bayesian models for data missing not at random in health examination surveys, arXiv:1610.03687.

Tolonen H., Aistrich A., Borodulin K. (2014) Increasing health examination survey participation by SMS reminders and flexible examination times. Scandinavian Journal of Public Health, 42(7): 712-717.

NBBC17 **Invited session 1. Digital epidemiology (June 21th, room 35.01.06, chair: Jaakko Nevalainen)**

TO BE ANNOUNCED

Krista Lagus

Aalto University

INFERRING HEALTH INFORMATION FROM NON-HEALTH SOURCES

Ingemar J. Cox[1,2]

[1] University College London, UK
[2] University of Copenhagen, Denmark

Collectively, people now create enormous quantities of digital data. Some is explicitly created, on, for example, social networks such as Twitter. Other data is unconsciously created as people interact with digital systems. For example, each user query to a web search engine is stored in a query search log, which records, amongst other things, the location of the query, the time and date of the query, and the words constituting the query.

While this data is not directly generated for health purposes, research has shown that it can be used for such. Examples include estimating the prevalence of influenza in a population, measuring the effectiveness of a vaccination campaign, and portmarket drug surveillance.

The advantages of using non-health sources depends on the circumstances, but can include (i) ease of data collection, (ii) timeliness, i.e. the lag between data creation, collection and analysis can be very short, (iii) the behavioural information inferred from the data is often unique or at least very difficult to acquire from alternative sources, and (iv) the number of participants is usually much greater than in traditional epidemiological studies. Digital data from non-health sources can complement traditional health data when it is harder to collect data in the physical world, or people have a difficulty reporting associations.

In this talk we will describe how digital data from non-health sources can be used for a variety of purposes related to health and medicine. The methods are based on statistical natural language processing and machine learning. A number of examples from our's and other's work will be given.

Keywords: computational epidemiology, statistical natural language processing, machine learning

References:

Eysenbach, G., (2006). Tracking flu-related searches on the Web for syndromic surveillance, *AMIA Annu Symp Proc.*, 244-248.

Lampos, V., Yom-Tov, E., Pebody, R., Cox, I.J., (2015). Assessing the impact of a health intervention via user-generated Internet content, *Data Mining and Knowledge Discovery*, *29*, 5, 1434-1457.

Yom-Tov, T., Gabrilovich, E (2013). Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries, *Journal Medical Internet Research*, *15*, (6):e124.

Inference Criteria for Hidden Markov Models in Biological Sequence Analysis

Broňa Brejová

Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava, Slovakia

Hidden Markov models (HMMs) are used for many tasks in DNA and protein analysis, including gene finding, sequence family modeling, and sequence alignment. In these application areas, we typically assume that the input DNA or protein sequence X was emitted by the model and seek the sequence of hidden states S which might have emitted X. Traditionally, such inference is done by the Viterbi algorithm, which finds sequence S maximizing the joint probability of X and S in the model. Alternatively, posterior decoding chooses the most likely state at every position of X separately, However, many other inference criteria can be formulated, often motivated by a particular application domain. I will discuss several examples of HMM inference criteria, with focus on modeling aspects and computational complexity. We have proved that some criteria lead to NP hard problems, while others can be computed by efficient algorithms.

HMMs were also recently introduced for modeling DNA sequencing data produced by Oxford Nanopore devices. This technology can produce very long sequencing reads but also suffers from high error rate. We propose to sample alternative sequences from the HMM to improve alignment of reads to a reference genome or to each other.

Keywords: Hidden Markov models, inference, computational complexity, sequence alignment.

References:

Brejová, B., Brown, D. G., and Vinař, T. (2007). The most probable annotation problem in HMMs and its application to bioinformatics. *Journal of Computer and System Sciences*, 73(7), 1060–1077.

Nánási, M., Vinař, T., and Brejová, B. (2010). The highest expected reward decoding for HMMs with application to recombination detection. In *Combinatorial Pattern Matching (CPM 2010)*, volume 6129 of *Lecture Notes in Computer Science*, pages 164–176. Springer.

Nánási, M., Vinař, T., and Brejová, B. (2015). Sequence annotation with HMMs: New problems and their complexity. *Information Processing Letters*, 115(6), 635–639.

Rabatin R, Brejová, B., and Vinař, T. (2016). Using Sequence Ensembles for Seeding Alignments of MinION Sequencing Data. arXiv preprint *arXiv:1606.08719*.

DETECTION OF ADAPTIVE SHIFTS ON PHYLOGENIES USING SHIFTED STOCHASTIC PROCESSES ON A TREE

Paul Bastide[123] and Mahendra Mariadassou[23] and Stéphane Robin[123]

[1] AgroParisTech, France
[2] INRA, France
[3] University Paris-Saclay, France

Comparative and evolutive ecologists are interested in the distribution of quantitative traits among related species. The classical framework for these distributions consists of a random process running along the branches of a phylogenetic tree relating the species. We consider shifts in the process parameters, which reveal fast adaptation to changes of ecological niches. We show that models with shifts are not identifiable in general. Constraining the models to be parsimonious in the number of shifts partially alleviates the problem but several evolutionary scenarios can still provide the same joint distribution for the extant species. We provide a recursive algorithm to enumerate all the equivalent scenarios and to count the number of effectively different scenarios. We introduce an incomplete-data framework and develop a maximum likelihood estimation procedure based on the EM algorithm. We also propose a model selection procedure, based on the cardinal of effective scenarios, to estimate the number of shifts and for which we prove an oracle inequality. Eventually, we will discuss the generalization of this approach to multivariate traits as weel as the case were the phylogenetic tree is not actually a tree.

Keywords: Random process on tree, Ornstein-Uhlenbeck process, Change-point detection, Adaptive shifts, Phylogeny, Model selection.

References:

Bastide, P., Mariadassou, M., and Robin, S. (2016). J. Royal Stat. Soc. B. DOI:10.1111/rssb.12206

MATHEMATICAL MODELLING AND BAYESIAN INFERENCE REVEALS NEW INSIGHTS INTO STRUCTURAL VARIATION OF THE HUMAN GENOME.

Mae Woods[1] and Chris Barnes[1]
[1] University College London

Structural variation in the human genome, in the form of deletion, insertion, inversion and translocation, occurs in both germline and somatic cells, and is observed frequently in cancer genomes. Recent accumulation in whole genome sequencing on paired tumor and non-tumor samples and the development of algorithms to estimate absolute copy number profiles, have led to a rich structural dataset from which we can further current inferences on the mutational landscape of the human genome.

It is known that the choice of repair pathway assigned to mend breaks in DNA affects the probability of structural variation. Hence, because these mutations are a direct consequence of the interplay between DNA damage and repair, the dependencies between the numbers of insertions and deletions observed might result in a deeper insight into activity of the DNA repair machinery, a group of processes that play a fundamental role in evolution.

We present a Bayesian approach to understanding the probability landscape of structural variation using approximately 2000 cancer genomes, comprising of 14 primary sites. First we hypothesize that similarities between cancers, which may be implicit by robustness of biological networks, could potentially be identifiable by analysis of structural measures. We show by statistical analysis on the data that there is a clear universal optimum in how much a chromosome can deviate from the size reported in the human reference genome. We find that after adjustment for confounding there are nontrivial dependencies between the probability of insertions, deletions and also translocations on a chromosome. This leads us to predict using a mathematical model that these highly constrained distributions on the length of a chromosome are the product of a dynamical system involving not only deletions and insertions, but also translocations. We show that this system appears to be common amongst cancer types, highlighting the importance of the information gained when multivariate correlations observed in big data are combined with the predictive power of theoretical models.

Keywords: Structural variation, Mutation processes, Evolution, Genetics.

FUNCTIONAL DATA MODELING OF DYNAMIC PET DATA

R. Todd Ogden[1], Yakuan Chen[1], and Jeff Goldsmith[1]

[1] Department of Biostatistics, Columbia University, New York, New York, USA

One major goal of dynamic positron emission tomography (PET) imaging, with particular relevance to the study of mental and neurological disorders, is the estimation of the spatial distribution of specific molecules throughout the brain. Current analysis strategies involve applying parametric models that require fairly strong assumptions, reducing information for each subject and each voxel/region into a single scalar-valued summary, and modeling each subject and each voxel/region sequentially. We will describe extensions of the analysis in three different directions: a nonparametric approach to the modeling of the observed PET data; a functional data analytic (FDA) approach to modeling the impulse response function; and the ability to consider observed PET data from multiple subjects in a single function-on-scalar regression model. We demonstrate the application of this approach and compare the results with those derived from standard parametric approaches.

Keywords: Functional data analysis, PET imaging.

ADVANCES IN LOCAL INFERENCE FOR FUNCTIONAL DATA WITH APPLICATION TO TONGUE PROFILE ANALYSIS.

Alessia Pini[1], Lorenzo Spreafico[2],
Simone Vantini[1], Alessandro Vietti[2]
[1] MOX - Department of Mathematics, Politecnico di Milano, Italy.
[2] ALPs - Alpine Laboratory of Phonetics and Phonology, Free University of Bozen-Bolzano, Italy.
Authors' names are alphabetically ordered

The talk will focus on the statistical comparison of ultrasound tongue profiles pertaining to different allophones pronounced by the same speaker (Vietti et al. 2015). Stimulated by this application we will introduce a general framework for *multi-aspect local non-parametric* null-hypothesis testing for functional data (Pini and Vantini 2016 and 2017, Pini et al. 2017). Sagittal tongue profile records can be modelled indeed as functions varying on a spatio-temporal domain. In detail: *multi-aspect* pertains to the fact the procedure allows the simultaneous investigation of different data features (i.e., aspects) like tongue vertical position, slope, concavity, velocity, and acceleration; *local* pertains instead to the fact the procedure can impute the aspect-specific rejection of the null hypothesis to aspect-specific regions of the spatio-temporal domain; finally, *non-parametric* refers to the fact that the procedure is permutation-based and it is thus finite-sample exact and consistent independently on data Gaussianity. For ease of clarity, the focus will be on the two-population test. Nevertheless, the approach is flexible enough to be adapted to more complex testing problems like functional ANOVA and functional linear regression.

Keywords: Functional data analysis, non-parametric inference, local inference, multi-aspect inference, articulatory phonetics.

References:

A. Pini and S. Vantini (2016), The interval testing procedure: a general framework for inference in functional data analysis. *Biometrics*, 72, 835–845.

A. Pini and S. Vantini (2017), Interval-Wise Testing for Functional Data, *Journal of Nonparametric Statistics*. To appear.

A. Pini, L. Spreafico, S. Vantini, A. Vietti (2017), Multi-aspect local inference for functional data: analysis of ultrasound tongue profiles. Tech. Rep. MOX, Dept. of Mathematics, Politecnico di Milano.

A. Vietti, L. Spreafico, and V. Galatà (2015), An ultrasound study of the phonetic allophony of Tyrolean /r/. *ICPhS 2015 Proceedings*.

MARKOV COMPONENT ANALYSIS AND FUNCTIONAL REGRESSION

Bo Markussen[1]
[1] Department of Mathematical Sciences, University of Copenhagen, Denmark

Decomposition of data into underlying components is one of the most used techniques in multivariate statistics and functional data analysis. *Principal Component Analysis (PCA)* is the foremost among the decomposition techniques. For a sample of multivariate functional data $X_n\colon [a,b] \to \mathbb{R}^d$ for $n = 1, \ldots, N$ functional PCA into q components is given by

$$X_n(t) = \sum_{k=1}^{q} S_{nk}\phi_k(t) + \text{error}.$$

The scores $S_{nk} \in \mathbb{R}$ can be interpreted as stochastically independent random variables (Tipping & Bishop, 1999), and the loadings $\phi_k\colon [a,b] \to \mathbb{R}^d$ are deterministic functions. In this talk we propose an alternative decomposition that we call *Markov Component Analysis (MCA)*. This is given by

$$X_n(t) = \sum_{k=1}^{q} Z_{nk}(t) + \text{error}.$$

Here the components $Z_{nk}\colon [a,b] \to \mathbb{R}^d$ are stochastically independent Markov processes. The covariance operators for Gaussian Markov processes can be parametrized by a factorizable structure that extends the concept of loadings from PCA. From a modelling point of view this structure is more versatile, while it retains many of the nice computational properties known from PCA.

In this talk we will compare MCA to PCA, present an algebra that allows for efficient computation within the MCA framework, and apply MCA to a functional regression problem of predicting horse lameness from three dimensional acceleration signals.

Keywords: Functional Data Analysis, Multivariate Analysis, Covariance estimation, Decomposition methods, Sparse representation, Functional Regression.

References:

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *J. R. Statist. Soc. B*, *61*, 611–622.

Sørensen, H., Tolver, A., Thomsen, M. H., and Andersen P. H. (2012). Quantification of symmetry for functional data with application to equine lameness classification. *J. Appl. Stat.*, *39*, 337–360.

CLIMATE TIME SERIES: SIGNAL, NOISE AND ADDITIVITY

Gudrun Brattström[1]
[1] Stockholm University, Sweden

Climatological time series (temperature, precipitation etc.) arise from three sources:

- Instrumental measurements
- Reconstructions from proxies, e.g. tree rings, ice cores, historical records...
- Computer simulations from climate models: 3D dynamical systems, in principle deterministic

In all three variability is assumed to be decomposable as climate signal + internal variability, the latter being "noise" produced by the internal dynamics of the system. The signal is the effect of external input into the system, forcings, e.g. volcanic eruptions, greenhouse gas emissions, solar activity (sunspots). Climate model simulations may be run using only one forcing, or perhaps a subset of known forcings. Climate scientists often use such simulations in an attempt to detect the effects of a particular forcing in the actual climate (instrumental or reconstructed). A common assumption in such studies is that the forcing effects in the simulations are additive, so that up to internal variability the effect of several forcings may be obtained from simulations using only one forcing at the time.

Is the additivity assumption true? I will discuss several different approaches to this question.

Keywords: Climate time series, Climate model simulation, Additivity, Forcing.

RECONSTRUCTING PAST LANDSCAPES – SPACE-TIME MODELLING OF COMPOSITIONAL DATA WITH ZERO COUNTS

Johan Lindström[1]
[1] Lund University, Sweden

Fossil pollen records can be used to understand regional interactions between human land use, past vegetation and climate. A study of fossil pollen covering 45 sites in Southern Sweden during 44 time periods from present to 7000 BC has recently been completed. From these pollen records the local proportions of five land cover types: coniferous and decidious forest, shrubs, open land and arable land (i.e. human land use), can be extracted. However, for climate modelling and to understand the evolution of human land use it would be desirable to interpolate the data into space-time maps.

Similar reconstructions have previously been done for 5 time periods over Europe using Gaussian Markov Random Fields (GMRFs) and Dirichlett distributions (Pirzamanbein, et. al, 2017). Here we extend the model in Pirzamanbein to handle: 1) spatio-temporal dependencies and 2) absence of some classes.

The Dirichlett model is suitable for the compositional pollen data since it automatically obeys the (0,1) and sum-to-one restrictions of the data. However, the Dirichlett distribution does not handle absence of classes from the distribution, i.e. observations with 0% probability of one class. To handle the zero-probabilites we propose a model that combines Dirichlett and Bernoulli observations of the same underlying GMRF, creating a joint model for presence/absence and proportions of the classes. Allowing us to create maps, and to model the advance of agriculture.

Keywords: Gaussian Markov Random Field, Dirichlett distribution, Fossil pollen, Southern Sweden, past vegetation

References:

Pirzamanbein, B., Lindström, J., Poska, A. and Gaillard, M-J. (2017). Modelling Spatial Compositional Data: Reconstructions of past land cover and uncertainties *arXiv*, 1511.06417v2

ADDRESSING THE STATISTICAL CHALLENGES OF ESTIMATING PAST CLIMATE FROM POLLEN SOURCES

James Sweeney[1] and John Haslett[2]

[1] School of Business, University College Dublin

[2] School of Computer Science & Statistics, Trinity College Dublin

Quantitative reconstructions of past climate have an intrinsic value as a source of insight into the Earth's climate history, including the timescales involved in abrupt changes in climate. Such reconstructions also provide a potentially valuable tool for evaluating the performance of general circulation models (GCMs), which are used to explore the potential future climatic consequences of anthropogenic changes to the Earth system.

In this talk we focus on inferring past climate from fossil pollen data, a "noisy" source of climate information obtained from lake bed sediment cores, and identify the statistical challenges which need to be overcome to harness pollen as a climate proxy. These are multiple in nature - the present day model training dataset, consisting of highly multivariate, zero-inflated compositional counts for vegetation, as well as measurements on several climate covariates including temperature and precipitation, presents numerous challenges of model choice and inference. In particular, we present a framework for modelling compositional count data subject to an excess of both zeroes and N's, a feature introduced to our pollen responses by the imposition of hierarchical structures. We also illustrate the drawbacks of omitting influential climate covariates in models, and the resulting impact on climate predictions. We conclude by providing historical climate predictions for a number of European locations, and attempt to evaluate the accuracy of these estimates.

Keywords: Statistical climatology, Zero and N-inflated compositional data, Inverse regression problems.

References:

Parnell, A.P., Sweeney, J., Doan, T.K., Salter-Townshend, M., Huntley, B., Allen, J.R.M. and Haslett, J. (2015). Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. *JRSS C*, *64 (1)*, (pp. 115–138).

Parnell, A.P., Haslett, J., Sweeney, J., Doan, T.K., Allen, J.R.M. and Huntley, B. (2017). Joint palaeoclimate reconstruction from pollen data via forward models and climate histories. *Quaternary Science Reviews*, *151*, (pp. 111–126).

Sweeney, J., Haslett, J. and Parnell, A.P. (2017). Zero & N-inflated binomial distributions with applications. *arXiv:1407.0064v5*.

JOINT MODELLING OF MULTIPLE LATENT PROCESSES AND A CLINICAL ENDPOINT: APPLICATION IN ALZHEIMER'S DISEASE

Cécile Proust-Lima

INSERM U1219, Université de Bordeaux

Alzheimer's disease, the most frequent dementia in the elderly, is characterized by multiple progressive impairments in the brain structure and in clinical functions such as cognitive functioning and functional disability. Until recently, these components were mostly studied independently while they are fundamentally inter-related in the degradation process towards dementia.

We propose a joint model to describe the dynamics of multiple correlated latent processes which represent various domains impaired in the Alzheimer's disease. Each process is measured by one or several markers, possibly non-Gaussian. Rather than considering the associated time to dementia as in standard joint models, we assume dementia diagnosis corresponds to the passing above a covariate-specific threshold of a pathological process modelled as a combination of the domain-specific latent processes. This definition captures the clinical complexity of dementia diagnosis but also benefits from an inference via Maximum Likelihood which does not suffer from the usual complications due to numerical approximations of multivariate integrals.

The method is illustrated on large French population-based cohorts of cerebral aging which include repeated information on brain structure (hippocampus volume, cortical signature) and/or clinical manifestations (cognitive functioning, physical dependency and depressive symptoms) as well as clinical-based diagnoses of Alzheimer's disease.

Keywords: aging, joint model, latent process model, multivariate longitudinal data

CAUSAL VALIDITY OF MARGINAL STRUCTURAL MODELS FOR EVENT HISTORY DATA AND CAUSAL LOCAL INDEPENDENCE GRAPHS

Kjetil Røysland[1] and Vanessa Didelez[2]

[1] University of Oslo, Norway
[2] University of Bremen, Germany

Survival analysis has become one of the fundamental fields of biostatistics. Such analyses are almost always subject to censoring. This necessitates special statistical techniques and forces statisticians to think more in terms of stochastic processes. The theory of stochastic integrals and martingales have therefore been important for the development of such techniques.

Causal inference has lately had a huge impact on how statistical analyses based on non-experimental data are done. The idea is to use data from a non-experimental scenario that could be subject to several spurious effects and then fit a model that would govern the frequencies we would have seen in a related hypothetical scenario where the spurious effects are eliminated. This opens up for using the Nordic health registries to answer new and more ambitious questions. However, there has not been so much focus on causal inference based time-to-event data or survival analysis.

The now well established theory of causal Bayesian networks is for instance not suitable for handling such processes. Motivated by causal inference event-history data from the health registries, we have introduced causal local independence models. We show that they offer a generalization of causal Bayesian networks that also enables us to carry out causal inference based on non-experimental data when there is continuous-time processes involved.

The main purpose of this work in collaboration with Vanessa Didelez, is to provide new tools for determining identifiability of causal effects of event history data that is subject to censoring. It builds on previous work on local independence graphs and delta-separation by Vanessa Didelez and previous work on causal inference for counting processes by Kjetil RÃ¸ysland.

We provide a new result that gives quite general graphical criteria for when causal validity of a local independence model is preserved in submodels. If the observable variables, or processes, form a causally valid submodel, then we can identify most relevant causal effects by re-weighting the actual observations. This is used to prove that the continuous time marginal structural models for event history analysis, based on martingale dynamics, are valid in a much more general context than what has been known previously.

SURVIVAL ANALYSIS AROUND A CROSS-SECTION AND UNOBSERVED HETEROGENEITY

Niels Keiding

University of Copenhagen, Denmark

Consider lifetimes originating at a series of calendar times $t_1, t_2, \ldots$. At a certain time t_0 a cross-sectional sample is taken, generating a sample of *current durations* (backward recurrence times) of survivors until t_0 and a *prevalent cohort study* consisting of survival times left-truncated at the current durations. A Lexis diagram is helpful in visualizing this situation.

Survival analysis based on current durations and prevalent cohort studies is now well-established as long as all covariates are observed.

The general problems with *unobserved covariates* have been well understood for ordinary prospective follow-up studies, with the good help of hazard rate models incorporating frailties: as for ordinary regression models, the added noise generates attenuation in the regression parameter estimates.

For current durations and prevalent cohort studies this attenuation remains, but in addition one needs to take account of the differential selection of the survivors from initiation t_i to cross-sectional sampling at t_0.

This talk intends to survey the recent development of these matters and the consequences for routine use of hazard rate models or accelerated failure time models in the many cases where unobserved heterogeneity may be an issue.

CONTROLLING FOR UNMEASURED CONFOUNDING IN TIME-TO-EVENT ANALYSES OF TREATMENT EFFECTS: THE MISSING CAUSE APPROACH

Michal Abrahamowicz[1] and Marie-Eve Beauchamp[2]
[1] McGill University, Canada
[2] Research Institute of the McGill University Health Centre, Canada

Unobserved confounding is the main source of bias in observational studies of drug safety and effectiveness. The instrumental variable (IV) approach, which uses physicians' prescribing preferences as an IV, was proposed to deal with this challenge and was shown to remove bias due to unobserved confounding, under certain assumptions. However, the IV method is limited to linear (risk difference) regression models, and its implementation in time-to-event analyses is difficult. We recently proposed an alternative 'missing cause' approach for the risk difference modeling of a treatment effect on a binary outcome (Abrahamowicz et al. 2016). We now extend this approach to time-to-event analyses, based on the Cox model, and evaluate it in simulations. Similar to the IV methodology, we assume that the treatment prescribed to a patient depends on both (1) the (observed and unobserved) patient's characteristics, and (2) the subjective prescribing preference of his physician (Brookhart et al. 2006). We then postulate that the discrepancy between the expected treatment of individual patients and treatment actually prescribed can be used as a marker for potential unobserved confounding. Based on these assumptions, we expand the model by including an interaction between the treatment indicator and the measure of treatment discrepancy, and use this model to obtain a corrected estimator of the treatment effect. In simulations, with strong unobserved confounding the proposed corrected estimator reduced the bias substantially and yielded lower mean squared error than the conventional model.

This work is related to STRATOS Survival Analysis group (TG8).

Keywords: Bias, Cohort Studies, Cox PH model, Pharmacoepidemiology, Simulations.

References:

Abrahamowicz, M., Bjerre, L.M., Beauchamp, M.-E., LeLorier, J., and Burne, R. (2016). The missing cause approach to unmeasured confounding in pharmacoepidemiology. *Stat. Med.*, *35*, 1001-1016.

Brookhart, M.A., Wang, P.S., Solomon, D.H., and Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiol.*, *17*, 268-275.

CAUSAL INFERENCE IN SURVIVAL ANALYSIS USING PSEUDO-OBSERVATIONS

Per Kragh Andersen[1] and Elisavet Syriopoulou[2] and Erik T. Parner[3]

[1] Section of Biostatistics, University of Copenhagen, Denmark
[2] Department of Health Sciences, University of Leicster, UK
[3] Department of Biostatistics, University of Aarhus, Denmark

For binary or quantitative outcomes *causal inference* can typically be performed using either the g-formula or using inverse probability of treatment weighting (IPTW) (e.g., Hernan and Robins, 2017). However, when dealing with a time-to-event outcome, special techniques for causal inference are required due to the inevitable presence of incomplete observation caused by right-censoring.

In this paper we will show how one, alternatively, may use so-called *pseudo-observations* (e.g., Andersen and Pohar Perme, 2010) when doing causal inference in survival analysis. The idea is to first transform the censored outcome data into pseudo-observations for the random variable of interest (e.g., if the t-year cumulative incidence corresponding to a given cause of failure is the parameter of interest then pseudo-observations for the t-year failure indicator for that cause are needed). Thereby, censoring is dealt with 'once and for all' and in the next step, the average causal effect may be estimated using the g-formula or IPTW as for a completely observed outcome.

We will show how this idea works and illustrate it both via Monte Carlo simulations and via an example from bone marrow transplantation in patients with acute leukemia.

Keywords: Average causal effect, Causal inference, Censored data, Pseudo-observations.

References:

Andersen, P.K. and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Stat. Meth. Med. Res.*, *19*,(pp. 71–99).

Andersen, P.K., Syriopoulou, E. and Parner, E.T. Causal inference in survival analysis using pseudo-observations *Statist. in Med.* (in press).

Hernan, M.A. and Robins, J. (2017). *Causal Inference*. Boca Raton: CRC/Chapman and Hall (in press).

CAUSAL INFERENCE IN SURVIVAL ANALYSIS – CONTINUOUS WEIGHS AND AN APPLICATION TO CANCER REGISTRIES

Pål Christie Ryalen and Kjetil Røysland[1]

[1] University of Oslo, Norway

Marginal structural models have served an important role for doing causal inference using non-experimental longitudinal data. An important example is in analysis of health registries. Using the marginal structural models, causal effects have been assessed by weighting observed data according to discrete time, inverse probability weights. Robins et.al. introduced Marginal structural models in discrete time. However, survival data is continuous and does not fit entirely in discrete time models. Røysland et. al. have considered continuous time likelihood ratio weights, that are theoretically correct in the survival setting.

We have developed software that estimate the continuous time weights, using the Aalen additive hazard model. Using this software we compare Nelson-Aalen estimates weighted according to the discrete and continuous weights. We also discuss consistency of the estimated weights and weighted Nelson-Aalen estimates.

We conclude with an application on prostate cancer using cancer registry data from Norway.

Keywords: Causal reasoning with survival data, Marginal structurel models, Aalen additive hazard model.

References:

Røysland, K. (2011) A martingale approach to continuous-time marginal structural models *Bernoulli 17*, 895–915.

Røysland, K. Didelez, V. Nygård, M. Lange, T. Aalen, O. Causal reasoning in survival analysis: Re-weighting and local independence graphs. Unpublished manuscript.

Identifying Interactions via Hierarchical Lasso Regularization

Ann-Sophie Buchardt[1]
[1] University of Copenhagen, Denmark

Penalised regression models such as the Lasso or elastic net are powerful methods which use sparsity to do feature selection in situations where the number of features measured is much larger than the number of observations. This may be the case for genomic data or multi-omics data, and when analysing such data we are concerned not only with identifying relevant features but also with identifying more complex relationships such as interactions, e.g. gene-gene or gene-environment interactions. However, it is not clear exactly how to consistently include features which are correlated and, in addition, interacting in penalised regression models.

We consider methods for identifying pairwise interactions in a linear regression model under different assumptions of hierarchy: No hierarchy allowing for all interactions being included at once; weak hierarchy allowing for the inclusion of only interactions between at least one pre-selected main effects and another main effect; strong hierarchy allowing for the inclusion of only interactions between pre-selected main effects. We motivate our approach by modelling pairwise interactions for quantitative variables and experiment with explicitly applying (and not applying) penalties on the main effects and interactions, thereby obtaining interpretable models. We compare the methods with existing approaches on simulated data using R.

Keywords: Hierarchical, Lasso, Interactions, Regression, Genetics.

STRATEGIES TO HANDLE MANDATORY COVARIATES USING MODEL- AND LIKELIHOOD-BASED BOOSTING

Riccardo De Bin[1]
[1] University of Oslo, Norway

Among the iterative methods exploited during recent years in statistical practice, particular attention has been focused on boosting. Originally developed in the machine learning community to handle classification problems, boosting has been successfully translated into the statistical field and extended to many statistical problems, including regression and survival analysis. In a parametric framework, the basic idea of boosting is to provide estimates of the parameters by updating their values iteratively: at each step, a weak estimator is fitted on a modified version of the data, with the goal of minimizing a loss function. Thanks to its resistance to overfitting, boosting is particularly useful in the construction of prediction models. Its iterative nature, moreover, allows straightforward adaptations to cope with high-dimensional data. In this talk, we first review and contrast two well-known boosting techniques, model-based boosting and likelihood-based boosting. We note that in the simple linear regression case they lead to the same results, provided there is a specific choice for their tuning parameters. This is not the case for more complex situations. As an example, we show the differences in survival analysis under the proportional hazards assumption. As a main contribution of the talk, we analyze strategies to include mandatory variables, i.e. those variables that for some reasons must enter in the final model, in a statistical model using the two boosting techniques. In particular, we examine solutions currently only considered for one and explore the possibility of extending them to the other. We show the importance of a good handling of mandatory variables in a biomedical study on colon cancer.

Keywords: boosting, data integration, high-dimensional data, prediction model, survival analysis

GENERALIZED PREDICTION OF RANDOM EFFECTS IN BALANCED AND UNBALANCED TWO-FACTORIAL RANDOM-EFFECTS MODELS

Razaw Al-Sarraj[1], Claudia von Brömssen[1] and Johannes Forkman[1]
[1] Swedish University of Agricultural Sciences

Linear random models, consisting of several unknown variance components, have been used extensively in many fields, for instance in the analysis of experiments in the agricultural sciences. In addition to several unknown variance components, random models comprise also random effects, which are predicted using the best linear unbiased predictor (BLUP). BLUPs are functions of the unknown variance components, usually estimated by known procedures, e.g. the maximum likelihood (ML) or the restricted maximum-likelihood (REML) procedures.

In certain situations, for example in small agricultural field experiments, non-positive estimates of variance components can be obtained, resulting in a difficulty to assess the precision in the BLUP. For inferential problems where frequentist methods fail to provide useful solutions, Tsui and Weerahandi (1989) introduced the generalized p-value for hypotheses testing, which later was followed by the generalized confidence intervals (Weerahandi, 1993).

In this study, generalized prediction intervals were derived for linear combinations of random effects. For both balanced and unbalanced data in two-way layouts, linear random-effects models were considered, with and without interaction. Coverage of generalized prediction intervals was estimated through simulation, based on an agricultural field experiment. Generalized prediction intervals were compared with prediction intervals derived using the REML method. Coverage of generalized prediction intervals was closer to the nominal value than coverage of prediction intervals using the REML procedure.

Keywords: Generalized prediction intervals, Random models, Random effects, REML.

References:

Tsui, K. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypothses in the presence of nuisance parameters. *J. Amer. Statist. Assoc.*, *84*, (pp. 602-607).

Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc.*, *88*, (pp. 899–905).

PREDICTING DERMAL ABSORPTION OF CHEMICALS FROM PHYSICO-CHEMICAL PROPERTIES AND EXPERIMENTAL CONDITIONS. A GENERALIZED LINEAR MIXED MODEL APPROACH.

Gilles Guillot

Risk Assessment and Scientific Assistance Department,
European Food Safety Authority, Parma, Italy.

Predicting the amount of chemical transported from the outer surface of the skin into the circulatory system is of interest for risk assessors, for example for the numerous pesticides under regulation for which no experimental data are available (EFSA, 2012). Towards this aim, we propose here a statistical approach based on generalized linear mixed model in which the fraction of chemical absorbed is explained by physico-chemical properties of the compound, the experimental conditions and some descriptors of the skin sample on which the chemical is applied.
We assume a Beta distributed response and a Bayesian setting. We computer posterior and predictive distributions together with model selection criteria using integrated nested Laplace approximations (Rue et al. 2009; Martins et al. 2013) with the R-INLA package. We train our model and explore its predictive ability on a data-set consisting of about 6500 skin samples and explore a family of sub-models containing up to thirteen explanatory variables.
We found that most variables considered explain a significant part of variation in dermal absorption. We contrast this with the need for simple and parsimonious models, in particular in a regulatory context.

Keywords: Quantitative structure-activity relationship model (QSAR), Beta regression, Integrated Nested Laplace Approximation (INLA).

References:

European Food Safety Authority (2012). Guidance on Dermal Absorption. *EFSA Journal 2012;10(4)* 2665.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series B (statistical methodology), 71(2)*, 319-392.

Martins, T. G., Simpson, D., Lindgren, F., Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis, 67*, 68-83.

Risk factors for treatment discontinuation for neovascular AMD

Inger Westborg, MD[1] and Aldana Rosso, PhD[2]

[1] Department of Clinical Sciences/Ophthalmology, Umeå University, Umeå, Sweden

[2] Centre of Registers South, Skåne University Hospital, Lund, Sweden, and Department of Radiology, Institution of Translational Medicine, Lund University, Malmö, Sweden

Neovascular age-related macular degeneration (nAMD) is one of the leading causes of visual impairment in the elderly population in the Western World. The standard treatment for patients with nAMD is anti-vascular endothelial growth factor (anti-VEGF) intravitreal injections. Clinical trials show that visual acuity improves with frequent visits and injections. However, approximately, between 30 - 60 % of the patients in clinical practice discontinue treatment already during the first year. Therefore is of paramount importance to identify which factors affect the treatment discontinuation.

We present a statistical analysis of the treatment given to 932 nAMD patients in the Skåne County diagnosed from 2013 through 2015. The data are obtained from the Swedish Macular Registry and the Skåne County Healthcare Database. We estimate the risk of terminating the treatment during the first year using a Poisson model. The model predictors are age, visual acuity, type of lesion, treatment drug, eye diseases and severe comorbidities present before and after nAMD diagnosis, and health-care provider. Multiple imputation is used to estimate missing values. We perform sensitivity analyses to estimate the variation of the model coefficients under different data conditions. In addition to the Poisson model, the data are analyzed using a classification tree in order to separate the patients into groups in which treatment discontinuation is more or less frequent.

Overall, we found that patients with low visual acuity and previous serious comorbidities at diagnostic treated at the university hospital have higher risk of terminating the nAMD treatment within a year. Patients treated with aflibercept have lower risk of early treatment termination compared to patients treated with ranibizumab. Our findings provide valuable information for treatment planning.

References:

Brown DM, Kaiser PK, Michels M, et al. Ranibizumab versus Verteporfin for Neovascular Age-Related Macular Degeneration. New England Journal of Medicine. 2006;355(14):1432-1444.

Evaluation of Quality of Care using Patient Follow-up Data

Michael Höhle[1,2] and Johannes Hengelbrock[1]
[1] Institute for Quality Assurance and Transparency in Healthcare, Germany
[2] Stockholm University, Sweden

The recently founded Institute for Quality Assurance and Transparency in Healthcare (IQTIG) is the central institution in Germany for the statutory quality assurance in health care. In accordance with its statutes, it is scientifically independent and works for, in particular, The Federal Joint Committee as well as the Federal Ministry of Health providing its expertise in various tasks of quality assurance of medical care.

One of the biostatistical tasks has been the development of new quality indicators for longitudinal data, where patient outcomes are tracked as subsequent operations after an initial surgery occur, i.e. so called patient follow-up data. We describe the development of two new indicator types for analysing such data within an survival context by taking censoring mechanisms of the data generating process into account: An un-adjusted Kaplan-Meier based survival rate indicator and a risk-adjusted standardized mortality ratio (SMR) based on the work of Breslow (1975). For each indicator type we discuss the construction of appropriate confidence intervals and how to embed the indicators into a sequential monitoring scheme applicable to the German hospital benchmarking performed at the IQTIG in the areas of cardiac pacemakers as well as hip and knee endoprosthesis. Results from a simulation study comparing the proposed methodology against other alternatives are also given.

Keywords: Statistics in Healthcare, Survival Analysis, Statistical Quality Control

References:

Breslow, N. (1975). *Analysis of Survival Data under the Proportional Hazards Model*. International Statistical Review, 43(1), 45–57.

Fay, M.P., Brittain, E.H. (2016) *Finite sample pointwise confidence intervals for a survival distribution with right-censored data*, Statistics in Medicine, 35(16):2726-40.

IQTIG (2017), *Ereigniszeitanalyse-Methodik für die Follow-up-Indikatoren nach QSKH-RL*, To become available from `http://www.iqtig.org` (in German).

THE IMPACT OF HIGH-SCHOOL COMPLETION - A MULTI-STATE MODEL FOR WORK, EDUCATION AND HEALTH IN YOUNG MEN

Rune Hoff[12], Ingrid S. Mehlum[1], Karina Corbett[1], Ferdinand Mohn[1], Petter Kristensen[1], Therese Hanvold[1] and Jon M. Gran[2]

[1] National Institute of Occupational Health, Norway
[2] Oslo Centre for Biostatistics and Epidemiology, Norway

Completing high school is associated with higher work participation, higher educational attainment and less health-related absence, even when adjusting for other known predictors. Few studies consider the dynamic nature of such outcomes over time. We assessed how high school completion affected work participation, education and health-related absence in young men, using multi-state models.

Baseline covariates and follow-up data on states of work, education and health-related absence were obtained from national registries for all males born in Norway between 1971 and 1976 (n=184 951). The impact of high school completion on transitions between states during age 21 to 35 was analysed using Cox proportional hazards models. The long-term effect of completion is illustrated by calculating state probabilities for all outcomes during follow-up. Population average effects were assessed using inverse probability weighting, re-weighting the observed data by individuals' inverse propensity of own high school completion status given their observed covariates. We compared results from univariate and reweighted analyses to illustrate how much of the effects were explained by known baseline confounders.

For Norwegian men, high school completion was strongly associated with higher work participation, more education and less health-related absence. When also considering important confounders, effects were reduced although still substantial.

Keywords: Multi-state models, health-related absence, work participation, high school completion.

A PSEUDO-OBSERVATIONS ESTIMATOR IN RELATIVE SURVIVAL ANALYSIS

Maja Pohar Perme[1] and Klemen Pavlič

[1] Faculty of Medicine, University of Ljubljana, Slovenia

Several estimators have been proposed for net survival estimation, but only recently, a consistent estimator has been introduced. Its use in practice has revealed an excessively large variance when estimating net survival of older age groups. We first simplify the problem by considering a non-censored case to show that the problem of large variance is intrinsic to the definition of net survival and not a property of a specific estimator. We then continue from the definition of net survival and generalize it to the censored case by the use of pseudo-observations. The estimator developed in this way has all the desired properties, we also provide a formula for its variance. Since pseudo-observations are available in several statistical packages, this new estimator is easy to implement. It has several interesting theoretical properties, its main advantage in practice is the fact that it does not require numerical integration. This also implies it can be directly used with life-table data, i.e. data grouped in intervals of time. We illustrate the properties of our proposal with simulations and a real data example of colon cancer patients. The research is related to the activities of the STRATOS initiative for developing guidance for the analyses of observational studies (www.stratos-initiative.org).

Keywords: Medical statistics, Survival analysis, Pseudo-observations, Relative survival.

SEMIPARAMETRIC MULTI-PARAMETER REGRESSION SURVIVAL MODELLING

Kevin Burke[1] and Frank Eriksson[2] and Christian Pipper[2]

[1] University of Limerick, Ireland
[2] University of Copenhagen, Denmark

We consider a log-linear model for survival data, where both the location and dispersion parameters depend on covariates and the baseline hazard function is completely unspecified. It is argued that this model provides the flexibility needed to capture many interesting features of survival data at a relatively low cost in model complexity.

Estimation procedures are developed based on identifying the counting process martingales. Moreover, asymptotic properties of the resulting estimators are derived using empirical process theory. Finally, a resampling procedure is suggested to estimate the limit distributions of the estimators. The finite sample properties of the estimators are investigated by simulation

Keywords: multi-parameter regression, semiparametric model, survival data

JOINT MODELLING ON EARLY NUTRITION AND ADVANCED β CELL AUTOIMMUNITY: A COMPARISON AMONG DIFFERENT APPROACHES

Essi Syrjälä[1], Jaakko Nevalainen[1], Jaakko Peltonen[2] and Suvi M. Virtanen[1,3]

[1] Faculty of Social Sciences, University of Tampere, Finland
[2] Faculty of Natural Sciences, University of Tampere, Finland
[3] National Institute for Health and Welfare, Helsinki, Finland

The incidence of Type 1 diabetes (T1D) has increased woldwide, with Finland having the highest incidence in the world. Contradictory evidence on the association between the food consumption and the development of β cell autoimmunity (pre-diabetes) or T1D exists but no specific dietary factor has yet been shown to be an unambiguous risk factor.

A prospective birth cohort of 6069 infants born in 1996-2004 with genetic susceptibility to T1D was recruited. Food records were frequently collected and diabetes-associated antibodies repeatedly measured. The endpoint of interest is repeated positivity for the antibodies and/or T1D.

We have used joint models to investigate the association between the food consumption and pre-diabetes. The idea is to couple a survival model with a suitable linear mixed effects model. This enables modelling of two phenomena at the same time efficiently and without bias. We have used three approaches for the joint models: basic joint models, joint models where the linear mixed effects model is replaced by a Gaussian process fit, and joint latent class mixed models. In my poster presentation, I am going to present a comparison between these three approaches.

Keywords: Joint model, Joint latent class mixed model, Linear mixed effects model, Gaussian processes, Early nutrition, Type 1 diabetes.

References:

Proust-Lima, C., Philipps, V., and Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *arXiv preprint arXiv:1503.00890*.

Rizopoulos, D. (2012). Joint models for longitudinal and time-to-event data: With applications in R. *CRC Press*.

Virtanen, S. M., Nevalainen, J., Kronberg-Kippila, C., Ahonen, S., Tapanainen, H., Uusitalo, L., Takkinen, H.-M., Niinistö, S., Ovaskainen, M.-L., Kenward, M. G., Veijola, R., Ilonen, J., Simell, O. and Knip, M. (2012). Food consumption and advanced β cell autoimmunity in young children with HLA-conferred susceptibility to type 1 diabetes: a nested case-control design. *The American journal of clinical nutrition*, *95(2)*, 471-478.

ASSESSING PRECISION IN COEFFICIENTS OF VARIATION BETWEEN AND WITHIN SUBJECTS USING GENERALIZED PIVOTAL QUANTITIES

Johannes Forkman[1]

[1] Swedish University of Agricultural Sciences, Sweden

In linear mixed-effects models with a single random-effects factor, there are two variance components: the random-effects variance, i.e., the inter-subject variance, and the residual error variance, i.e., the intra-subject variance. The intra- and inter-subject coefficients of variation are the square roots of the corresponding variances divided by the mean. In many applications, it is practice to report variance components as coefficients of variation. For example, in bioanalytical method validation, coefficients of variation specify within- and between-run precision. In agricultural field experiments, coefficients of variation are often used to identify experiments with large field variation.

Weerahandi (1993) introduced generalized confidence intervals for variance components in linear mixed models. Generalized confidence intervals are approximate, but may perform well for practical purposes. The present study proposes generalized confidence intervals for intra- and inter-subject coefficients of variation. Coverage of confidence intervals was investigated through simulation. The proposed generalized confidence intervals were compared with, and found to perform better than, confidence intervals for coefficients of variation based on variance-stabilizing transformation (Shoukri, 2006).

Keywords: Bioanalytical method validation, Generalized confidence interval, Linear mixed model, Split-plot experiment.

References:

Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Statist. Assoc. 88*, (pp. 899–905).

Shoukri, M.M., Elkum, N., and Walter, S.D (2006). Interval estimation and optimal design for the within-subject coefficient of variation for continuous and binary variables. *BMC Med. Res. Methodol., 6*, (pp. 1–10).

DATA ANALYSIS USING HIERARCHICAL GENERALIZED LINEAR MODELS WITH R

Youngjo Lee[1], Lars Rönnegård[2] and Maengseok Noh[3]
[1] Seoul National University, Korea
[2] Dalarna University, Sweden
[3] Pukyong National University, Korea

Since their introduction, hierarchical generalized linear models (HGLMs) have proven useful in various fields by allowing random effects in regression models. Interest in the topic has grown, and various practical analytical tools have been developed. We have summarized developments within the field and, using data examples, show how to analyse various kinds of data using R. The work is currently being published as a monograph. It provides a likelihood approach to advanced statistical modelling including generalized linear models with random effects, survival analysis and frailty models, multivariate HGLMs, factor and structural equation models, robust modelling of random effects, models including penalty and variable selection and hypothesis testing. This example-driven book is aimed primarily at researchers and graduate students, who wish to perform data modelling beyond the frequentist framework, and especially for those searching for a bridge between Bayesian and frequentist statistics.

Keywords: Hierarchical likelihood, Hierarchical Generalized Linear Models, Statistical Modelling.

NBBC17 Contributed session 5 Generalized linear mixed models and R (June 20th, room 35.3.12, chair: Helle Sørensen)

TMB: A FLEXIBLE FRAMEWORK FOR DEVELOPING MIXED MODEL SOFTWARE IN R

Kasper Kristensen[2] Geir Drage Berentsen[1] Mollie Brooks[2]
Anders Nielsen[2] Hans J. Skaug[1]

[1] University of Bergen, Norway.[2] National Institute of Aquatic Resources, Technical University of Denmark, Denmark

Mixed models are commonly used in biometry. TMB (https://github.com/kaskr/adcomp) is an R package that allows quick and flexible implementation of mixed model R packages. The flexibility of TMB comes from invoking C++, so the user of TMB must master some level of C++ programming. In this respect TMB is similar to Rcpp, but it also has functionality for integrating over complex configurations of random effects via the Laplace approximations. The new R package glmmTMB for fitting overdispersed and zero-inflated mixed models has been developed this way, and is a major example of the approach that we advocate. We also give examples of other projects in which TMB has been used to implement mixed models.

Keywords: Mixed models, C++, Laplace approximation, R packages.

COMPARISON OF HIDDEN MARKOV CHAIN MODELS AND HIDDEN MARKOV RANDOM FIELD MODELS IN ESTIMATION OF COMPUTED TOMOGRAPHY IMAGES

Kristi Kuljus[1], Fekadu L. Bayisa[2], David Bolin[3], Jüri Lember[1] and Jun Yu[2]

[1] University of Tartu, Estonia
[2] Umeå University, Sweden
[3] Chalmers University of Technology and University of Gothenburg, Sweden

There is an interest to replace computed tomography (CT) images with magnetic resonance (MR) images for a number of diagnostic and therapeutic workflows. We explore the problem of predicting CT images from a number of magnetic resonance imaging (MRI) sequences using regression approach. Two principal areas of application for estimated CT images are dose calculations in MRI based radiotherapy treatment planning and attenuation correction for positron emission tomography (PET)/MRI. The main purpose of this work is to compare the performance of hidden Markov (chain) models (HMMs) and hidden Markov random field (HMRF) models for predicting CT images of head. The study shows that HMMs have advantages over HMRF models in this particular application. Obtained results suggest that HMMs deserve a further study for investigating their potential in modeling applications where the most natural theoretical choice would be the class of HMRF models.

Keywords: Computed tomography, Magnetic resonance imaging, Pseudo-CT, Hidden Markov model, Hidden Markov random field, Radiotherapy, Attenuation correction.

MULTI-RESOLUTION CLUSTERING OF TIME DEPENDENT FUNCTIONAL DATA WITH APPLICATIONS TO CLIMATE RECONSTRUCTION

Konrad Abramowicz[1], Sara Sjöstedt de Luna[1], and Johan Strandberg[1]

[1] University of Umeå, Sweden

A multi-resolution clustering approach is presented to detect latent groups from observed dependent functional data. Given a lattice of (time) points, a function is observed at each grid point. We assume that the latent (unobservable) groups vary slowly over time (the lattice). We consider the case when at different time scales (resolutions) different groupings arise, with groups being characterized by distinct frequencies of the observed function-types. We propose and discuss a non-parametric double clustering based method, which identifies latent groups at different scales, based on bagging Voronoi strategies. We present an application of the introduced methodology to the annual seasonal patterns of varved lake sediment data, aiming at reconstructing winter climate regimes in northern Sweden at different resolutions during the last six thousand years.

Keywords: functional clustering, multi-resolution, climate reconstruction

THE SURPRISING IMPLICATIONS OF FAMILIAL ASSOCIATION IN DISEASE RISK

Morten Valberg[1], Mats J. Stensrud[1] and Odd O. Aalen[1]

[1]Dept. of Biostatistics, Oslo Center for Biostatistics and Epidemiology, University of Oslo, Norway

A wide range of diseases show some degree of clustering in families. The family history is therefore important when clinicians make risk predictions. A familial aggregation is often quantified in terms of a familial relative risk (FRR). Even if this measure may seem simple and intuitive at first glance -as an average risk prediction- its implications are not straightforward.

We use two statistical models for the distribution of risk of disease in a population: A dichotomous risk model that gives an intuitive understanding of the implication of a given FRR, and a continuous risk model that facilitates a more detailed computation of the inequalities in disease risk. Published estimates of FRRs are used to produce Lorenz curves and Gini indices that quantifies the inequalities in risk for a range of diseases.

We demonstrate that even a very moderate familial association in disease risk implies a very large difference in risk between individuals in the population. We give examples of diseases where this is likely to be true, and we further demonstrate the relation between the point estimates of FRRs and the distributions of risks in the population.

There is an ongoing debate about the role of chance in cancer development, which was fueled by a Science paper by Tomasetti and Vogelstein. They claimed that two thirds of cancers are due to 'bad luck'. Cancer of the small intestine was deemed the second 'most random' of the cancers studied by Tomasetti and Vogelstein. We show that this cancer is likely to have a risk distribution that is as skewed as the distribution of wealth in the world. This points towards determinants other than randomness being much more important for the development of the cancer.

Keywords: Cancer, Familial relative risk, Inequality, Lorenz curve.

References:

Tomasetti, C., and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, *347(6217)*, (pp. 78–81).

SOME STATISTICAL CHALLENGES ON THE ROAD FROM GWAS TO PERSONALIZED RISK PREDICTION

Krista Fischer[1] and Kristi Läll[2]
[1] Estonian Genome Center, University of Tartu, Estonia
[2] Institute of Mathematics and Statistics, University of Tartu, Estonia

We will discuss some statistical issues encountered in the process of development and validation of algoriths for personalized prediction of complex disease risk, based on Genetic Risk Scores (GRS) as well as other known environmental and lifestyle-related risk factors. Usually the GRS is defined as a linear combination of effect allele counts of several Single Nucleotide Polymorphisms (SNPs), whereas the SNPs and their corresponding weights are based on results of a large-scale meta-analysis of Genome-Wide Association Study (GWAS).

As large GWAS meta-analyses are often based on all available genotyped cohorts, it is possible that the validation cohort has been included in the discovery study. We demonstrate that even if the cohort forms no more than 1-2% of the total meta-analysis sample, one could get dramatically misleading conclusions while validating the GRS. That is especially true in cases where the GRS accounts for the effects of thousands of SNPs – shown to be most efficient for several complex diseases, incl Type 2 Diabetes (e.g Läll et al 2016). Also, one ideally needs two validation samples – one that is used to compare different versions of the GRS and select the optimal one, and another one for final validation of the GRS. We discuss options for the optimal sample selection and compare alternative scenarios.

Next we demonstrate how the effects of phenotypic risk factors and GRS are combined to an overall risk score and illustrate how the absolute risks can be calculated in the case of Type 2 Diabetes risk estimation. The methodological discussion is illustrated with examples based on the Estonian Biobank cohort.

Keywords: Genetic risk scores. Biobank-based data analysis. Risk prediction algorithms.

References:

Läll, K., Mägi R., Morris A., Metspalu A., Fischer K. (2017) Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine, 19(3)*, (pp 322-329).

A RANK-BASED BAYESIAN APPROACH TO COMBINING GENOMIC STUDIES

Valeria Vitelli[1], Elja Arjas[1,2], Thomas Fleischer[3], Vessela N. Kristensen[1,3], Magne Thoresen[1], Manuela Zucknick[1] and Arnoldo Frigessi[1,4]

[1] University of Oslo, Norway
[2] University of Helsinki, Finland
[3] Oslo University Hospital Radiumhospitalet, Norway
[4] Oslo University Hospital, Norway

In a typical meta-analysis in the context of genomic studies, different gene lists arise from different studies or platforms, and the aim is to combine them in a unifying consensus gene list to give further insight to the biological interpretation. Another frequent situation in applications is the availability of microarray data from different sources (platforms, laboratories, studies), which are collected to the same purpose: being able to combine them would enhance the power of the methods and thus strengthen the biological conclusion. However, due to the heterogeneity of data sources, preprocessing steps are needed to make the data comparable prior to the analysis: this might affect the analysis reproducibility, and even its results. Hence, it would be beneficial to avoid those preprocessing steps by using a method which is insensitive to measurements heterogeneity.

We tackle these problems via a rank-based approach. The use of ranks in combining genomic studies is relevant for the biological question, since we are often interested in the most (or least) expressed genes for a given pathology. Moreover, ranks are insensitive to heterogeneity in the measurement scales, thus allowing to combine information across studies and platforms without preprocessing: indeed, most normalization strategies for genomic data are based on ranks, or on transformations which do not affect the relative magnitude of the signal across genes. Last but not least, ranks are more robust to outliers and measurement error, both often quite heterogeneous across studies / platforms. For combining genomic data, we propose to use a Bayesian Mallows model for ranks, first introduced in Vitelli *et al.*, 2017. Due to its Bayesian nature, the method easily handles missing data by augmentation procedures, which are embedded in the estimation process, and allows to also quantify all uncertainties associated to the final result.

Keywords: Genomics, Bayesian methods, Data integration, Meta-analysis.

References:

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A. and Arjas, E. (2017). Probabilistic preference learning with the Mallows rank model. *arXiv preprint 1405.7945*.

NBBC17 Contributed session 7 Genomic studies (June 19th, room 35.3.12, chair: Claus Ekstrøm)

A Novel Variant Recalibration Method for Detection of Rare Mutations in Next-Generation Sequencing Experiments

Sarunas Germanas

Vilnius University, Lithuania

Next-generation sequencing (NGS) is often used to identify genetic variants. This sequencing technique suffers from large sequencing errors and demands sophisticated mathematical methods to control this problem. Therefore probability of variant detection depends on the variant caller. According to the best variant indentification practices (Mckenna A. et al., 2010) one of important phases of variant detection is variant recalibration. Standard tools use reference datasets (for example, 1000 Genomes project) to recalibrate called variants in order to raise the sensitivity and specificity of the caller. In our work we focus on rare mutations which usually are not present in open reference datasets. Therefore we do not use any other genetic dataset and use only probabilities of genotypes and positions of suspected variants. Proposed variant recalibration procedure consists of two stages - division of target region regarding the estimated genotype probabilities and recalibration! of genotype probabilities. For division of the target region we use change point detection methods (non-parametric likelihood test, (Avery P. J. et al., 1999)). After recalibration procedure we use mixture model to divide mutated and not mutated positions in a subregion. We apply offered and known methods to 1000 genomes data and check sensitivity and specificity using positons with known mutations status.

Keywords: Genetics, variant calling, variant recalibration, NGS, mutation.

References:

Mckenna A. et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*.

Avery P. J., Henderson D. A. (1999). *Detecting a changed segment in DNA sequences*. *Applied Statistics.*

ESTIMATION OF PAIRWISE RELATEDNESS AND JOINT IBD

Elizabeth Thompson[1] and Bowen Wang[1]
[1] University of Washington, Seattle, WA, USA

Genetic marker data provide evidence of the segments of genome that individuals share identical by descent (IBD) from a single DNA copy in a recent common ancestor. Inference of such IBD segments provide measures of both the global (genome-wide) and location-specific (local) realized relatedness among individuals. Global relatedness is used in studies of heritability of traits, and of population structure and history. Local relatedness is used in the mapping of causal DNA underlying traits of medical or economic importance.

There are a number of well-known methods for estimation of pairwise realized relatedness. However, many current methods do not allow for inbreeding, do not adjust for allelic associations (linkage disequilibrium), and do not take into consideration that genomes descend in large segments generation to generation. We present methods that model or allow for these factors and can provide more accurate estimators of the precise segments of genome shared IBD.

Although for many purposes estimates of pairwise IBD suffice, in other cases patterns of joint IBD among individuals can be informative as to trait etiology. A particular case arises for rare variant alleles within a single functional gene, giving rise to similar phenotypes. Although IBD estimation jointly among larger sets of individuals remains a challenge, current methods permit analyses of trios of individuals which can provide significant gains in inference.

Keywords: Eastimation of Relatedness; local gene identity by descent (IBD); joint IBD inference; Rare variants.

References:

Brown, M.D., Glazner, C.G., Zheng, C. and Thompson, E. A. (2012). Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics 190*, 1447–1460.

Wang, B., Sverdlov, S. and Thompson E.A. (2017). Efficient estimation of pairwise relatedness. *Genetics 205*, 1063–1078.

ESTIMATING NATURAL DIRECT EFFECTS IN RCTS WITH MULTIPLE TIME-VARYING INTERMEDIATES

Aksel Karl Georg Jensen[1] and Theis Lange[1]

[1] University of Copenhagen, Denmark

In randomized clinical trials with time-to-event outcomes e.g., death, the foremost interest is to study the overall or total effect of an intervention on survival. Often however, e.g., via regular patient visits, multiple intermediate variables are measured several times during follow-up. In such settings with multiple time-varying intermediates, many trials offer an abundance of information about underlying mechanisms that can shed light on not just *if* the treatment has an overall effect but also *how* and *through which* paths the treatment affects the outcome.

We present two approaches that can estimate natural direct effects without introducing any models for the various mediators. Consequently, we cannot quantify the effect that is mediated via a specific mediator. But via the natural direct effect, we can quantify how much of the total effect we can and cannot account for via the measured intermediates. Specifically, both approaches need to address two challenges that inherently exist when modelling survival in a counterfactual setting with multiple time-varying intermediates: (1) Some individuals who survive beyond time t would have died before time t had they been treated differently. (2) Some individuals who die before time t would have survived beyond time t had they been treated differently. Notably, we do not know what values the intermediates would have taken after the observed time of death had the individual survived. We use recursively defined weights to undertake a sufficient correction at each time point where intermediates are measured.

The first approach focus on a discrete setup based on the prespecified timepoints where post randomization measures are registered. Using a stepwise approach, we estimate a natural direct effect: the relative chance of survival beyond a specific timepoint. Following ideas presented in Vansteelandt, Bekaert and Lange (2012), the second approach rely on models for the survival time and avoid models for the intermediates. The key idea here, is to treat the unobserved counterfactual survival times as missing values that can be imputed via parametric survival models such as Weibull models.

Keywords: Mediation analysis, natural direct effects, RCT, Survival.

References:

Vansteelandt, S., Bekaert, M. and Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods.,1(1):Artic.7*.

INTERVENTION TIME SERIES ANALYSIS WITH ITS APPLICATION TO THE SWEDISH TOBACCO QUITLINE

Xingwu Zhou[1], Alessio Crippa[1], Rosaria Galanti[1] and Nicola Orsini[1]

[1] Karolinska Institutet, Sweden

[2] Centre of Epidemiology and Community Medicine, Sweden

Knowledge about the series of phone calls received by a smoking cessation quitline in response to different interventions aiming at reducing tobacco smoking is currently lacking. Aim of this study is to examine the possible effect of four types of policies on the calling rates to the Swedish smoking cessation quitline: a campaign on passive smoking (Jan 2001); placing larger warnings on cigarette packs (Sept 2002); banning smoking from restaurants (Jun 2005); and a 10% tax increase (Jan 2012). We used 16-years of monthly data collected between January 1999 to December 2014 (192 months) counting a total of 162,978 phone calls.

Upon definition of four pre-post intervention intervals, we used intervention time series ARIMA (Auto-Regressive Integrated Moving Average) models (Box and Tiao, 1975) where the outcome was defined as calling rates expressed per 100,000 smokers. Rate ratio (RR) at 6 months after intervention together with a 95% confidence interval (CI) were derived from the model. Within an overall decreasing trend in the population of smokers in Sweden, we were able to detect differential effects of smoking policies on the calling rates to the quitline, the most effective being the campaign on passive smoking and the larger warnings signs on the cigarette packs.

Keywords: Intervention time series, smoking quitline, ARIMA, ARIMAX

References:

Box, G.E.P., and Tiao, G. C. (1975). "Intervention analysis with applications to economic and environmental problems". *Journal of the American Statistical Association. 70*(349), 70–79.

PERSONALIZED COMPUTER SIMULATIONS OF BREAST CANCER TUMORS TREATED WITH NEOADJUVANT CHEMOTHERAPY WITH AND WITHOUT BEVACIZUMAB: A PROOF-OF-CONCEPT

Xiaoran Lai[1], Alvaro Köhn-Luque[1], Oliver M. Geier[2], Øystein Garred[2], Thomas Fleischer[2], Valeria Vitelli[1], Manuela Zucknick[1], Therese Seierstad[2], Elin Borgen[2], Anne-Lise Børresen-Dale[2], Vessela N. Kristesen[2], Olav Engebråten[2] and Arnoldo Frigessi[1]

[1] University of Oslo, Norway
[2] Oslo University Hospital, Norway

In the *Neo-Ava* clinical trial, 131 patients with HER2 negative primary tumors were randomised to receive chemotherapy in a neoadjuvant setting with or without anti-angiogenic therapy over 24 weeks. Predicting the response to systemic treatment in these breast cancer patients is highly complicated. Multiple, interacting biological processes operate at different spatial and temporal scales often vary from patient to patient.

To better understand the dynamic effect of chemotherapy in patients' tumor during the full treatment period, we formulate a multi-scale mathematical model that couples continuous and discrete systems. We show here the full pipeline of data collection, model personalization, initialization and validation for both responding and non-responding patients of the Neo-Ava cohort. We also show simulations of alternative schedules that predict improved treatment outcomes.

Our mathematical model quantitatively predicts the effect of specific drug dosages and schedules in patients of the NeoAva trial. By leveraging the mathematical model, we can also identify relevant parameters and mechanisms for treatment response, which might assist in the design of future clinical trials.

Keywords: Mathematical model, Multiscale modelling, Cancer treatment, Drug effect.

References:

T. Alarcón, H. M. Byrne, and P. K. Maini. A multiple scale model for tumor growth. *Multiscale Modeling Simulation*, 32:*440*, 2010.

M. R. Owen, I. J. Stamper, M. Muthana, G. W. Richardson, J. Dobson, C. E. Lewis, and H. M. Byrne. Mathematical modeling predicts synergistic antitumor effects of combining a macrophage-based, hypoxia-targeted gene therapy with chemotherapy. *Cancer research*, *718 : 2826 - 2837* , 2011.

ESTIMATING PARTIAL CORRELATION WITH DATA MISSING NOT AT RANDOM

Tetiana Gorbach1[1] and Xavier de Luna[1]
[1] Department of Statistics, USBE, Umeå University, Sweden

The partial correlation is a common measure of association between two variables of interest while removing the effect of control variables. The motivating application for this work is a study of the relation between longitudinal changes in brain structure and cognition after adjusting, e.g., for the influence of age on both brain and cognition changes. Thus, partial correlation between changes in brain (fMRI measures) and cognition (memory test measures) are computed to investigate a change-change association. Dropout is common in longitudinal fMRI studies, and the resulting missing data mechanism depends typically not only on observed data but also on the missing observations (missing not at random situation). Therefore, complete cases analyses or analyses based on a missing at random assumption will provide biased results. In order to allow for missing not at random data, we deduce uncertainty intervals for a partial correlation. The intervals are computed for a given nominal coverage probability reflecting both sampling variability and the uncertainty added by the fact that the missing data may be missing not at random. The results are justified theoretically and through simulation experiments.

Keywords: non-ignorable dropout, sensitivity analysis, longitudinal change-change study, uncertainty interval, fMRI, cognition studies.

References:

Genbäck, M., Stanghellini, E., de Luna, X. (2014). Uncertainty intervals for regression parameters with non-ignorable missingness in the outcome. *Stat. Pap.*, *56*,(pp. 829–847).doi:10.1007/s00362-014-0610-x

Gorbach, T., Pudas, S., Lundquist, A., Orädd, G., Josefsson, M., Salami, A., de Luna, X., Nyberg, L. (2016). Longitudinal association between hippocampus atrophy and episodic-memory decline. *Neurobiol. Aging*, *51*,(pp. 167–176) doi:10.1016/j.neurobiolaging.2016.12.002.

Vansteelandt, S., Goetghebeur, E., Kenward, M., Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sin.*, *16*,(pp. 953–979).

On the Linfoot Informational Correlation

Andreas Kryger Jensen[1], Rikke Berntsen[2] and Jacob B. Hjelmborg[2]

[1] University of Copenhagen, Denmark
[2] University of Southern Denmark, Odense, Denmark

The Linfoot informational correlation satisfies among very few the Renyi criteria making it highly interesting for studying independence and functional dependence between random variables. Based on the fundamental mutual information of random variables it bridges information theory and statistics.

The mutual information $I(X,Y)$ of a pair of random variables (X,Y) taking values in $\mathbb{R}^2$ with probability densities absolutely continuous with respect to the Lebesgue measure is defined as the quantity

$$I(X,Y) = \mathbb{E}_{X,Y}[\log \frac{f_{XY}(X,Y)}{f_X(X)f_Y(Y)}]$$

where f_{XY} is the joint density function of (X,Y) and f_X and f_Y are the marginal density functions. This highly celebrated functional is equivalent to the Kullback-Leibler measure of divergence of distributions from independence and is a cornerstone in information theory with the capacity of revealing very general and complex associations - at least when estimation is feasible. In 1957 E.H. Linfoot introduced the injective transform $I(X,Y) \mapsto \sqrt{1-\exp(-2I(X,Y))}$ of the mutual information into the unit interval. The Linfoot informational correlation is zero if and only independence is obtained and is one if information is conveyed functionally between the random variables. Further, it takes the absolute value of the Pearson product moment correlation in case of Gaussian random variables.

We aim to make the Linfoot transform operational in certain bivariate settings, in particularly when classical correlation measures fall short. The bivariate copula density function determines the Linfoot correlation and we compare different non-parametric estimators for the empirical copula density function and for that of classical archimedian copulas, eg. the Clayton-Oakes copula. We present an application on twin data.

Keywords: Mutual Information, Linfoot Correlation, Empirical Copula, Bivariate Twin Data.

References:

Linfoot, E. H. (1957). An informational measure of correlation. *Information and Control 1*, 85–89.

NBBC17 Contributed session 9 Statistical methodology (June 20th, room 35.3.12, chair: Esben Budtz-Jørgensen)

Methods for quantifying fruit dehiscence

Jan-Eric Englund
Swedish University of Agricultural Sciences

To quantify the time until a fruit breaks in a mixer mill, the random impact test (RIT) followed by a calculation of the halftime to dehiscence is one method used. There are some different alternatives to do the analysis, but it seems that some of the recommendations have used a model that is not relevant for this type of data. Here we discuss different solutions and the problems with some of the solutions already used.

The basic idea with the random impact test is to take closed fruits (for example, $n = 20$ fruits) and do the following procedure:

- Put them in a mixer mill, agitate for 5 seconds and count the number of open fruits.
- Replace the intact fruits in the mill and agitate for another 5 seconds, count the number of open fruits.
- Replace the intact fruits in the mill and repeat the procedure with agitation times 10, 20, 40 seconds (the observed time-points are linear on the logarithmic scale).

The recommendation in some references is now to make a logistic regression based on the cumulative proportion of open fruits versus the logarithm of the time-points (Lenser & Theissen, 2014). However, this model has some problems:

1) The same fruits are used repeatedly throughout the experiment causing dependences not considered in the model.
2) It does not use the fact that the observed times are censored.
3) The time-points where the cumulative proportions of open fruits are 0% or 100% are removed and not included in the calculations.

Here we use a model for the random impact test described by a failure time model with the log-normal distribution (or alternatively the log-logistic distribution). This model can handle the three drawbacks described above.

The halftime to dehiscence is calculated from the failure time distribution and to do the estimation initial estimates for the parameters μ and σ in the normal distribution are needed. It turns out that a very simple way to calculate these initial estimates give a surprisingly good and competitive estimate of the halftime to dehiscence.

Keywords: Logistic regression, Failure time distribution, Random impact test.

References:

Lenser, T. and Theißen, G. (2014). Quantifying Fruit Dehiscence Using the Random Impact Test (RIT). *Bio-protocol 4(15)*.