NBBC09

2nd Nordic-Baltic Biometric Conference

10-12 June 2009 - Tartu, Estonia

Conference Book

Contents

NBBC09: conference program

Tuesday, 9th June
14:00 – 18:00 Pre-conference course (day 1)

Wednesday, 10th June

9:00 – 13:00	Pre-conference course (day 2)		
14:00 – 14:30	**Opening plenary**, chair: Krista Fischer – L111 Kristjan Haller (Vice-Rector of the University of Tartu), Geir Egil Eide (President of the IBS Nordic Region)		
14:30 – 15:30	**Keynote lecture:** Nanny Wermuth, *What does indirect confounding imply for meta-analyses?*		
15:30 – 16:00	Coffee break		
16:00 – 17:30	I1: **Statistics in demography** (chair: Elja Arjas) L111 16:00 Ene-Margit Tiit, *Census – the first step in the history of (bio-) statistics* 16:35 Juha Alho, *Dual System Estimation of Drug Abuse* 17:10 Discussion (discussant Esa Läärä)	C1: **Bioinformatics I** (chair: Daniel Gudbjartsson) L402 16:00 Jukka Corander, *Bayesian clustering and feature selection for cancer tissue samples* 16:22 Jukka Sirén, *Reconstructing Population Histories from Single-Nucleotide Polymorphism Data* 16:44 Jüri Lember, *Measuring goodness of segmentation performed with hidden Markov models* 17:06 Rimantas Eidukevičius, *A Simple Stochastic Cell Cycle Model with Application to Cytometry Data*	C2: **Spatio-temporal modeling** (chair: Eric Renshaw) L202 16:00 Juha Heikkinen, *Assessing the uncertainty of the polygonal declustering estimator of a spatial mean* 16:22 Laura Saltyte, *Spatial-temporal modeling of Baltic Sea coastal zone parameters* 16:44 Ottmar Cronie, *Improved estimation in a spatio-temporal growth-interaction model* 17:06 Jaan Liira, *European biodiversity is still impacted by nature – spatial autocorrelation in mixed models*
18:30 – 21:00	**Welcome reception at the University History Museum, Toome Hill**		

Thursday, 11th June

9:00 – 10:00	**Keynote lecture** (chair: Bendix Carstensen) Martyn Plummer, *Bayesian hierarchical modelling with JAGS*		
10:00 – 10:30	Coffee break		
10:30 – 12:00	I2: **Bioinformatics** (chair: Jüri Lember) L111 10:30 Timo Koski, *De Novo Detection of DNA: Regulatory Binding Regions Using a Bayesian Random Frame Model* 11:05 Daniel Gudbjartsson, *Learning from the results of genome-wide association studies* 11:40 Discussion (discussant Jukka Corander)	C3: **Causal inference** (chair: Carlo Berzuini) L202 10:30 Olli Saarela, *Bayesian nonparametric monotonic regression* 10:52 Elja Arjas, *Predictive Bayesian inference and dynamic treatment regimes: the MACS data revisited* 11:14 Peter Jakobs, *Major treatment incompliance (including unplanned "cross-over" of study treatment): structural nested failure time model as an alternative analysis method?* 11:36 Krista Fischer, *A principal stratification-based model to estimate the effect of blinding in a clinical trial with open and blind arms*	C4: **Statistics in epidemiology** (chair: Juha Karvanen) L122 10:30 Elina Parviainen, *Feature extraction in visualizing and describing a predictive classifier: A case study* 10:52 Jurate Saltyte Benth, *Modelling and prediction of weekly influenza A specimens in England and Wales* 11:14 Nora Fenske, *Boosting Additive Quantile Regression for Investigating Childhood Malnutrition* 11:36 John Öhrvik, *Factor Analysis of the Metabolic Syndrome Identifies two Factors with Different Survival Patterns in Elderly*
12:00 – 13:30	Lunch at restaurant *Püssirohukelder*, Lossi 28		

13:30 – 15:00	I3: **Causal inference in genetics** (chair: Jukka Corander) L111 13:30 Stijn Vansteelandt, *Semiparametric tests for sufficient cause interactions* 14:05 Carlo Berzuini, *Causal inference in genetic epidemiology* 14:40 Discussion (discussant Claus Ekstrøm)	C5: **Statistics in ecology and forestry** (chair: Jaan Liira) L122 13:30 Birgir Hrafnkelsson, *Estimation of discharge rating curves with B-splines* 13:52 Mark Brewer, *A Temporal Compositional Analysis of Water Quality Monitoring Data* 14:14 Artur Nilson, *Test of the Family of r^{th}-functions for Growth and Distribution Models* 14:36 Crispin Mutshinda, *Teasing out the workings of community dynamics*	C6: **Clinical tests and measurement** (chair: Kalev Pärna) L202 13:30 Rima Kregzdyte, *Statistical detection of cut-off cadmium concentration in breast cancer etiopathogenesis study* 13:52 Rossana Moroni, *Statistical modelling of measurement errors in gas chromatographic analyses of blood alcohol content* 14:14 Bendix Carstensen, *Practical aspects of assessing agreement of clinical measurement methods* 14:36 Kristi Kuljus, *Comparing Experimental Designs for Benchmark Dose Calculations for Continuous Endpoints*
15:00 – 15:30	**Poster session with coffee** (L111 and Foyer)		
15:30 – 17:00	I4:**Life course epidemiology** (chair: Thor Aspelund) L111 15:30 Mervi Eerola, *Comparing methods for life-course analysis – an example: pathways to adulthood* 16:00 Sven Ove Samuelsen, *Case-cohort and nested case-control studies: Differences and similarities* 16:30 Kristiina Rajaleid, *Size at birth, adult body mass index, and risk of myocardial infarction in the SHEEP study*	C7: **Bioinformatics** II (chair: Jukka Sirén) L122 15:30 Marijus Radavičius, *Local reverse-complement symmetry of DNA sequences* 15:52 Tomas Rekašius, *A method of visualization of DNA sequences* 16:14 Erinija Pranckeviciene, *Elucidating predictably different phenotypes by multiclass classification and clustering*	C8: **Statistical modeling** (chair: Mark Brewer) L202 15:30 Søren Højsgaard, *Where are the cows and what are they doing* 15:52 Lauri Jauhiainen, *Effect of a pen in group feeding trials: modeling data from suckler cows* 16:14 Johannes Forkman, *The performance of best linear unbiased prediction in small randomised complete block experiments* 16:36 Peter Dalgaard, *Ideas about likelihood-based data analysis in R*
17:15 – 18:45	**Guided walks in Tartu** (see the social programme section for options)		
19:30 – 22:30	**Conference dinner at the restaurant "Atlantis"**		

Friday, 12th June

9:00 – 10:00	**Keynote lecture:** (chair: Jun Yu) L111 Eric Renshaw, *Stochastic Growth/Interaction Strategies for Simulating Spatial-Temporal Marked Point Processes*		
10:00 – 10:30	Coffee break		
10:30 – 12:00	I5: **Spatial statistics** (chair: Juha Heikkinen) L111 10:30 Geir Aamodt, *Cluster of disease: making or solving problems* 11:05 Aki Vehtari, *Advances in Gaussian processes for spatial epidemiology* 11:40 Discussion (discussant Jun Yu)	C9: **Statistical genetics** (chair: Tanel Kaart) L122 10:30 Matti Pirinen, *Bayesian QTL mapping based on reconstruction of recent genetic histories* 10:52 Bob O'Hara, *Estimating heritability of fluctuating asymmetry in Sticklebacks* 11:14 Klara Verbyla, *Comparison of Bayesian models for genomic selection using real dairy data*	C10 **Survival analysis** (chair: Peter Dalgaard) L202 10:30 Karri Seppä, *Mean and median survival times of cancer patients should account for informative censoring and general mortality predictions* 10:52 Juha Karvanen, *Visualizing covariates in proportional hazards model* 11:14 Priyantha Wijayatunga, *Asymptotic properties of collective conditional likelihood estimators for Bayesian network classifiers with censored data* 11:36 Szilard Nemes, *Bootstrap confidence intervals for dynamic path models*
12:00 – 12:30	**Closing** – L111		

Poster program

The posters will be put up on the back wall of Auditorium 111 and in the Foyer of Liivi 2. The presenting author will be available at the poster during the second coffee-break on Thursday (15:00–15:30).
Posters:

P1 Yaqiong Cui, *Bayesian Predictive Classifiers*

P2 Jouni Hartikainen, *Spatio-Temporal Analysis of Disease Incidence with Sparse Gaussian Processes*

P3 Søren Højsgaard, *Reproducible statistical analysis with StatWeave*

P4 Maris Hordo, *Climatic signals extracted from ring-width chronologies of Scots pine from Estonia*

P5 Timo Hurme, *Estimating location and variability parameters from classified potato tuber size data*

P6 Marika Kaakinen, *Life course analysis on the effects of FTO on the adulthood BMI in the Northern Finland Birth Cohort 1966*

P7 Ülle Kirsimägi, *Prognostic factors for short- and long-term graft survival in kidney transplantation in Estonia*

P8 Diana Laarmann, *Estimating Tree Survival on the Estonian Forest Research Plots Network*

P9 Pekka Marttinen, *Bayesian modeling of recombination events in bacterial populations*

P10 Szilard Nemes, *Bias in Odds Ratios by Logistic Regression Modelling and Sample Size*

P11 Janek Saluse, *Income-related inequality in the utilisation of health care in Estonia*

P12 Teija Seppänen, *Survival analysis of rectal cancer patients in Finland using parametric mixture models*

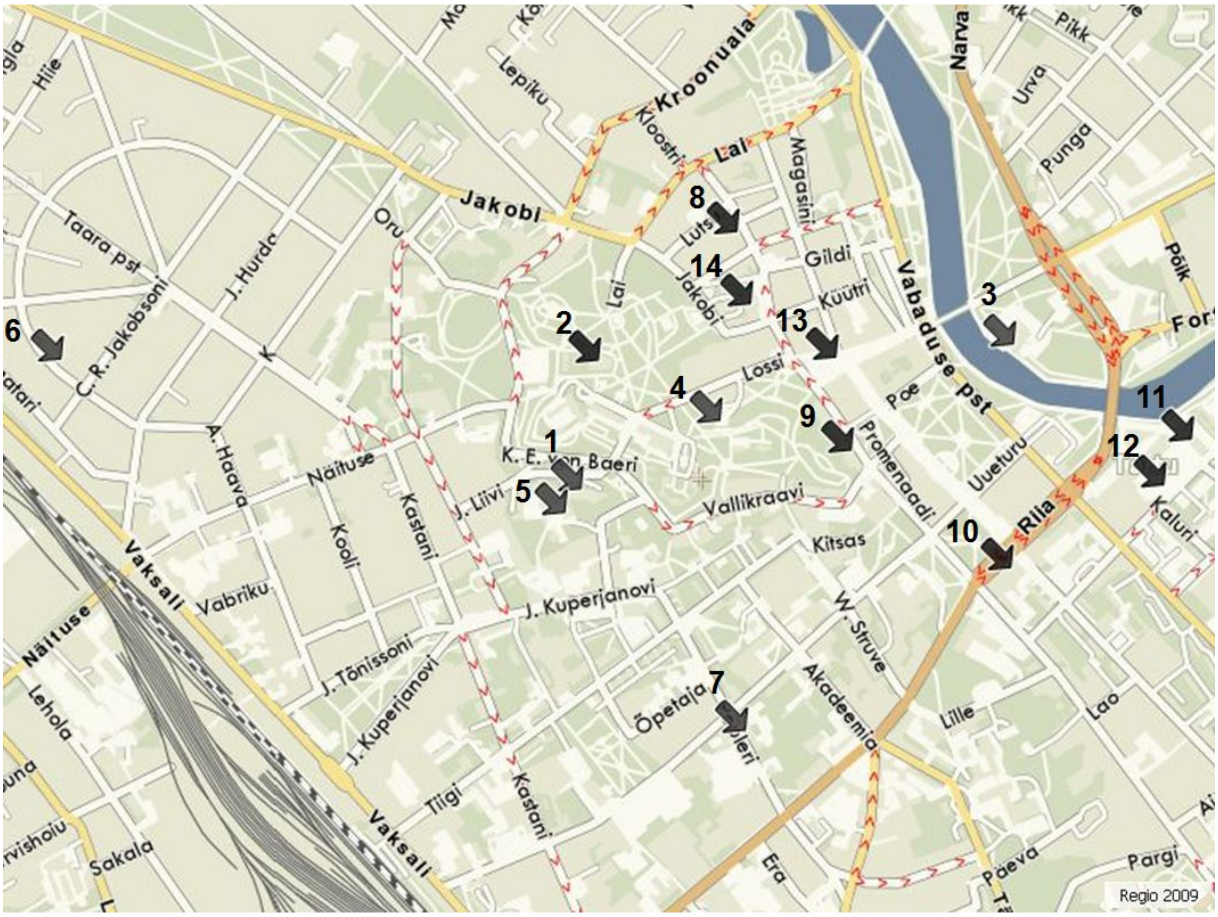
P13 Tiina Seppänen, *Analysis of overdiagnosis occurring in prostate cancer incidence in Finland*

P14 Allan Sims, *ForMIS - Forest Modeling Information System*

P15 Arvo Tullus, *Ordination of floristic data from young deciduous forest plantations using Non-metric Multidimensional Scaling (NMDS)*

P16 Jarno Vanhatalo, *Spatial variations in alcohol related deceases in Finland: Case study with sparse Gaussian process models*

Map of Tartu city center with conference venue, main hotels and locations for social events



1. Conference venue (Liivi 2)
2. University History Museum (Welcome reception on Wednesday)
3. Restaurant *Atlantis* (Conference Dinner on Thursday)
4. Restaurant *Püssirohukelder* (Lunch on Thursday)
5. Hotel Park
6. Hostel Vikerkaare
7. Hostel Pepleri
8. Guesthouse Tampere maja
9. Hotel Barclay
10. Hotel Pallas
11. Hotel Dorpat
12. Bus station
13. Town hall square
14. Main building of the University of Tartu

Welcome

Nordic Region and the Baltic Group of the International Biometric Society welcome you to the second Nordic-Baltic Biometric Conference. After the first very successful joint conference in Foulum, 2007, the Nordic Region and Baltic Group have decided to continue the tradition of joint conferences. This year the conference is an historical event, being the first ever biometrics conference in the Baltic region.

The conference has brought together more than 100 delegates from a wide geographical area. We are particularly happy to see a large number of Baltic participants – this is a clear sign of the increasing popularity of biometric research in this area.

The scientific program covers many diverse areas of biometric research. The programme committee has worked hard to put together the invited programme with 3 distinctive keynote talks and 15 invited speakers. In addition to that, there are 38 contributed talks and 15 posters representing exciting research results from a wide spectrum of applications, ranging from the analysis of DNA patterns to spatial modeling of environmental data.

The Local Organizing Committee has done some hard work putting the detailed time-schedule together, booking venues for the conference and social events. Our experience shows that with nowadays technology the local organizing committee can actually be international, with geographical distances creating no major complications in the work – our Skype-meetings lasted often several hours. We are grateful to the Institute of Mathematical Statistics and the Faculty of Mathematics of the University of Tartu for providing the great venue for the Conference in the heart of Tartu.

We hope the conference encourages research cooperation between Nordic and Baltic scientists, and the social programme provides opportunities for informal discussions and fun.

Tere tulemast Tartusse!

Welcome to Tartu!

Jukka Corander
on behalf of the Scientific Programme Committee

Krista Fischer
on behalf of the local Organizing Committee

Program Committee

Jukka Corander (PC chair, Åbo Akademi University, Turku, Finland)
Geir Aamodt (Norwegian Institute of Public Health, Oslo, Norway)
Irina Arhipova (Latvian University of Agriculture, Jelgava, Latvia)
Thor Aspelund (Icelandic Heart Association, Reykjavik, Iceland; University of Iceland, Reykjavik, Iceland)
Bendix Carstensen (Steno Diabetes Center, Gentofte, Denmark)
Marijus Radavicius (Institute of Mathematics and Informatics, Vilnius, Lithuania)
Jun Yu (Swedish University of Agricultural Sciences, Umeå, Sweden)
Esa Läärä (University of Oulu, Finland)
Krista Fischer (LOC chair, MRC Biostatistics Unit, Cambridge, UK; University of Tartu, Estonia)

Local Organizing Committee

Krista Fischer (LOC chair, MRC Biostatistics Unit, Cambridge, UK; University of Tartu, Estonia)
Heti Pisarev (Conference secretary, Department of Public Health, University of Tartu, Estonia)
Inge Ringmets (Department of Public Health, University of Tartu, Estonia)
Mare Vähi (Institute of Mathematical Statistics, University of Tartu, Estonia)
Märt Möls (Institute of Mathematical Statistics, University of Tartu, Estonia)
Tõnu Kollo (Institute of Mathematical Statistics, University of Tartu, Estonia)
Ene Käärik (Institute of Mathematical Statistics, University of Tartu, Estonia)
Andres Kiviste (Estonian University of Life Sciences, Tartu, Estonia)

Conference Host

NBBC09 is hosted by two institutes of the University of Tartu: Department of Public Health and Institute of Mathematical Statistics. The researchers at the Department of Public Health are actively involved in various research projects on epidemiology and Public Health in Estonia, the topics including investigation of HIV prevalence and risk factors in high-risk populations, longitudinal surveys on adolescent health behavior, studies on health services use and accessibility in population, and many more. The statisticians at the Department of Public Health provide statistical support to these studies and are also responsible for all biostatistics and epidemiology teaching and most of the statistical consulting for the Faculty of Medicine.

The Institute of Mathematical Statistics is the only academic unit in Estonia which educates statisticians. Being responsible for teaching statistics at the University of Tartu the Institute is also very active in statistical research. The main research area has traditionally been multivariate statistics (under the lead of Professors Ene-Margit Tiit, Tõnu Möls, Tõnu Kollo, and Kalev Pärna). In recent years the research areas have been extended to several exciting new directions including financial mathematics, survey sampling, machine learning, Bayesian analysis and bioinformatics. The Institute of Mathematical Statistics has throughout the years been actively involved in biostatistical consulting, supporting the research of life scientists in Estonia. In addition to the academic work, the Institute has played an important role in promoting the good use of statistics in society and supporting the development of official statistics in Estonia. Professor Ene-Margit Tiit is probably well known to most of Estonians through her clear and educative articles often appearing in the main newspapers. She has also been actively involved in the conduct of population censuses and other important studies on demography in Estonia.

Speaker/Chair Information

The location of your session is shown in this Conference Book. Please be on time for your session, check in with the session chair, and test the A/V equipment. The time allocated for your presentation is 50 minutes (+ 10min for discussion) for keynote speakers, 30 minutes (+5) for invited speakers and 20 (+2) minutes for contributed speakers.

Audio/Visual Equipment

Every room is equipped with a PC connected to a LCD projector. The computers contain up-to-date software for the main presentation formats (PowerPoint, MS Office, PDF) and have USB connections for memory cards. Please transfer your presentation onto the desktop. Overhead transparency projectors are also available at request. Please make sure to arrive at your session at least ten minutes before its scheduled start. Before the session begins, all presenters should set up and test their presentation and the connection with the LCD projector.

Session Chairs

The role of the chair is to ensure the smooth execution of the session. Make sure to:

- Contact the speakers before the session, to verify who will present and make sure there are no technical problems with their presentation.
- Begin the session on time.
- Ensure that the presentations, including questions, do not overstep their time frame.
- Keep presentations in the order shown in the program (starting at the stated time, even in cases where the previous speaker has not been present), to allow participants to jump between sessions.
- Introduce the speaker and the title of each presentation.
- Express visually to the speaker how many minutes (5, 1) are left, using either your hands or prepared cards.
- At the end of each presentation ask for questions and thank the speaker.

Poster presentations

The posters can be put up on Wednesday (in the auditorium L-111 and Foyer, to the indicated spaces). The presenters are expected to be present at their poster during the second coffee break on Thursday, from 15:00 to 15:30.

General information

Registration/information desk

The registration/information desk will be near the main auditorium (L-111) and cafeteria and will be open Tuesday (13:00–14:00), Wednesday (12:00–15:00), Thursday and Friday during coffee breaks.
At registration you will receive your badge and other conference materials. We recommend picking up your registration material as soon as you arrive.

Messages

A message board will be located in the Foyer of Liivi 2, next to the main auditorium L-111.

Coffee Breaks

Coffee and refreshments will be served in the Foyer of Liivi 2.

Lunches

On Thursday, the lunch will be served at the restaurant “Püssirohukelder” (“Gun-powder cellar”) from 12:00–13:30. There will be snacks and refreshments available for a light lunch (also possible to take away) after the closing on Friday at the conference venue.

Internet

There is free WIFI access at the conference venue and most of the hotels, restaurants and cafeterias in Tartu. In addition you may use the PC-s connected to the internet at the basement of Liivi 2 (003).

Banks and money

The currency in Estonia is Estonian Kroon (EEK), with the exchange rate of 1 EUR = 15.64 EEK. There are many ATMs in the Tartu city center and they all accept all major European credit and debit cards. The cards are similarly accepted in most shops and restaurants, even for quite small transactions (however, most taxi drivers in Tartu only accept cash).

Transportation from/to Tallinn

The bus trip from Tallinn to Tartu takes approximately 2hrs 30min. There have been buses organized by the conference departing on June 9th and June 10th from the Tallinn Airport and ferry terminals and back to Tallinn on June 12th, for the participants who requested that. On Friday, there will be buses taking people back to Tallinn Airport and Ferry terminals, departing from the conference venue at 13:30 (one-way bust-trip costing 270 EEK = 17 EUR).
It is also possible to take a regular bus from the Tallinn Bus station (about 5-10min taxi drive from the Airport, 15-20min drive from the ferry terminals, taxi trip costing about 60-100EEK, bus trip about 150 EEK). There is a bus from Tallinn to Tartu (and from Tartu to Tallinn) in every 30 minutes until 21:00 (and less frequently until 23:00).

Social programme

All social events are included in the registration fee. Please contact the registration desk if you wish to purchase additional tickets for accompanying persons.

Welcome reception

The NBBC09 hosts welcome you to the buffet reception at the University History Museum at Toome Hill on Wednesday, 18:30 – 21:00 (see map)

Guided city walks on Thursday

On Thursday after the scientific programme finishes, there will be a choice of guided city walks available. For the first-time visitors of Tartu, the classic walk will take you to the main historical sights in the town center. For people who have been in Tartu before (and possibly followed a guided tour before), there will be the possibility to explore the interesting area called *Supilinn*, with old wooden buildings and a special atmosphere and see the beautiful botanical gardens of the University.
All walks start at the conference venue at 17:15 and will finish between 18:30-18:45.

Dinner

The Conference Dinner will be served at the restaurant “Atlantis” – on the bank of the river Emajõgi, just over the pedestrian bridge from the town hall square (see map). You will have the opportunity for informal chat with other participants, while enjoying a delicious meal and a musical performance.
The dinner starts at 19:30 and will finish approximately 22:30.

List of abstracts

I Keynote talks

What does indirect confounding imply for meta-analyses?

Nanny Wermuth

Chalmers/Gothenburg University, Sweden

When analysing data, we all look out for possible interactive effects and possible common unobserved explanatory variables, that is for direct confounders which may fully or partially generate the dependences we see.

If we are in the fortunate situation of getting data from a randomized intervention study, such as a controlled clinical trial, we may even feel save that no severe distortions of dependences will occur. However, we also know that repeated controlled clinical trials may lead to inconclusive results, such as those that had been reported for trials with blood-thinning medication given to patients that have experienced a stroke.

For this particular situation, an explanation is that the studied patient groups may have been be inhomogeneous regarding their status. If the blood vessels are clogged, the medication will help, if they had already burst, the medication will harm. Formally, this is an interactive effect with an unobserved variable that may be present even if all direct effects of unobserved variables on the treatment variable have been removed by randomized allocation of individuals to treatments.

Suppose now, that in a properly randomized study, where there are no direct confounders, there are also no such inhomogeneities, and we condition on all possible, directly and indirectly explanatory variables, then there still be severe distortions in observed dependences when some of the variables of a generating process are not observed. Such types of distortions have recently been characterized for simple types of generating processes and named indirect confounding.

It is discussed how one can identify the possible presence and the absence of such distortions by using graphical representations of the generating processes and which effects these types of distortions may have on meta-analyses.

Bayesian hierarchical modelling with JAGS

Martyn Plummer

International Agency for Research on Cancer, France

JAGS is Just Another Gibbs Sampler. It is a program for the analysis of Bayesian graphical models using Gibbs Sampling, closely based on the program BUGS.

There are two primary motivations for JAGS. The first is to have an extensible BUGS "engine", allowing users to analyze non-standard problems without having to write their own software from scratch. A second motivation is to provide a platform for exploring topics in Bayesian modelling. This will be illustrated with a critical appraisal of the deviance information criterion for model choice, which has been popularised by the WinBUGS package.

Keywords: MCMC, deviance information criterion, BUGS.

Stochastic Growth/Interaction Strategies for Simulating Spatial-Temporal Marked Point Processes

Eric Renshaw

Department of Statistics and Modelling Science, University of Strathclyde, United Kingdom

The construction and analysis of spatial-temporal marked point processes has been fuelled by two separate fields of study. In biology, plants are often affected by others that compete with them for nutrient and natural resources. Whilst fundamental to the study of porous and granular materials is the modelling and statistical analysis of random systems of hard particles. A general high-intensity packing algorithm is presented that covers both situations in order to infer properties and generating mechanisms of space-time stochastic processes. Marks $m_i(t)$ $(i = 1, \dots, n)$ have location x_i and change size through the deterministic incremental size change

$$m_i(t+dt) = m_i(t) + f(m_i(t))dt + \sum_{j \neq i} h(m_i(t), m_j(t); \|x_i - x_j\|)dt \ . \qquad (1)$$

Here $f(\cdot)$ denotes the mark growth function and $h(\cdot)$ is an appropriate spatial interaction function; random variation is easily induced, e.g. via the simple immigration-birth-death process. If $m_i(t+dt) \leq 0$ then point i dies.

Different forms for the growth and spatial interaction functions produce different types of spatial structure. Here we consider variants of the power-law logistic process for $f(\cdot)$ and hard- and soft-core functions for $h(\cdot)$ that play a central role in the study of both forest and insect growth. Parameters may be estimated by using a maximum pseudo-likelihood technique for patterns that are sampled at a *fixed* time point, and a least squares procedure for those measured at *successive* time points. Following a brief historical presentation of spatial-temporal analysis, the technique is demonstrated through application in forest settings.

Now whilst the admission of immigration, birth and death injects sufficient randomness into the system for the algorithm to work well in biological scenarios, in materials science no such random process is likely to be present. This means that the technique is now wholly deterministic and so the stochastic nature of the process has to be viewed afresh. Six different possible approaches are therefore presented from fully stochastic to deterministic, and are first illustrated by applying them to the simple immigration-death process. They are based on: (a) exact event-time pairs; (b) time-increments; (c) tau-leaping; (d) Langevin-leaping; (e) chemical reaction rates; and (f) deterministic reaction rates. Remarkably little attention has been paid to the relationships which exist between these techniques, and this area of investigation holds many potentially exciting future challenges.

Applying these ideas to marked point processes requires two specific extensions. First, $f(\cdot)$ and $h(\cdot)$ have to be decomposed into general stochastic birth and death components. Second, as each marked point is affected differently we have to employ an individual-based approach. The current procedure, namely a combination of (1) and a stochastic immigration-death process, may be generalised to encompass the other approaches (a)-(e) across a wide range of disciplines. Moreover, it is relatively easy to extend the process by allowing particles to move under interaction pressure, and hence squeeze through small gaps from areas with high intensity to areas of low intensity, in order to produce maximally packed patterns. If the exact algorithm (a) incurs too large a compute-time penalty, then the approximations (b)-(e) should be analyzed in sequence in order to assess the trade-off between pattern structure and computational efficiency. This is of particular importance when developing models for large structures or systems under high-intensity packing which may take a long time to burn-in. Studies are currently being undertaken to generate models that accurately replicate three-dimensional packing structures for mixed-sized particle systems which previously could only be simulated by using ‘force-biased’ and ‘collective rearrangement strategies’. Whilst a further promising avenue would be to transfer methods recently developed for chemical reaction systems with a low to moderate number of molecules across to marked point processes. The scope for future development in this arena is enormous.

Keywords: Growth-interaction processes, maximum packing, simulation, spatial-temporal pattern.

References:

Renshaw, E. (2009) Spatial-temporal marked point processes: a spectrum of stochastic models. *Environmetrics* (to appear).

Renshaw, E. (2010) *Stochastic Population Processes*. Oxford University Press (to appear).

Renshaw, E. & Comas, C. (2009) Space-time generation of high-intensity patterns using growth-interaction processes. *Statistics and Computing* (to appear).

Renshaw, E, Comas, C. & Mateu, J. (2008) Analysis of forest thinning strategies through the development of space-time growth-interaction simulation models. *Journal of Stochastic Environmental Research and Risk Assessment*, **23**, 275-288.

Renshaw, E. & Särkkä, A. (2001) Gibbs point processes for studying the development of spatial-temporal stochastic processes. *Computational Statistics and Data Analysis*, **36**, 85-105.

Särkkä, A. & Renshaw, E. (2006) The analysis of marked point patterns evolving through space and time. *Computational Statistics and Data Analysis*, **51**, 1698-1718.

II Invited talks

I1: Statistics in demography

Census – the first step in the history of (bio-) statistics

Ene-Margit Tiit

Institute of mathematical statistics, University of Tartu, Estonia

The first censuses, being the first steps in statistics, were organized more than 5000 years ago in Babylon. Hence we can say that statistics belongs to the distinguished company of the oldest sciences in the world. As the censuses have the aim to give information for population statistics, from here it follows that population statistics and also biostatistics are very old sciences. From here we can conclude that this area of knowledge has been very important for people in all times. Also we see that the importance of population statistics has been acknowledged from very ancient time. When speaking about the results expected from census data we can differentiate two most important and specific tasks:

- Getting description of a population at the critical moment of a census – to make a so-called snapshot of the population.
- Getting information for making forecasts of the population, to foresee the situation of the population and it's most important characteristics in future.

Both these tasks have offered big challenges for statisticians in all times.

The first step contains all problems connected with data collection, cleaning, processing and quality assurance. Today it is difficult to imagine how it is possible to process information from millions records without programmed computers. But there are more curious facts in the history of censuses: in ancient Peru the numbers were marked with different knots, as the society was illiterate.

The next step – forecasting – belongs to the core of scientific philosophy. In fact, population projections are the very first mathematical models for forecasting random processes. The early physical and mechanical models (e.g. Newton law) were comparatively simple as they described non-random processes. The idea to use mathematical models for forecasting the population growth was a qualitatively new approach in the history of science.

It is quite interesting to follow the reflections of all these processes here, in the Nordic corner of the world. Of course, at the time when the first censuses were carried on in big cities of Babylon or Egypt, our few ancestors were nomads or lived in caves and primitive huts. But during the last centuries also Nordic countries and University of Tartu particularly have added their contribution into garner of bio- and population statistics.

We were not among the first, who started to make censuses, but we have made some important steps in developing the data handling and forecasting methodology.

When speaking about collection of population data we can see very different ideas to register all inhabitants and to keep such registers actual. In some areas of Russian Empire all people had to register their passports annually; in all totalitarian regimes there were quite detailed lists of inhabitants. But the only sustainable registers of population have been created as a system of registers (books) of members of congregations in Sweden by vicars, and so in Sweden, but also in other Nordic countries – Norway and Finland – exist oldest and the best population registers in the world. We, Estonians know this system by its ruins, created here more than two centuries ago, but repeatedly destroyed by alternating occupiers.

When speaking about building statistical models useable in population and biostatistics, we have to look also the University of Tartu (established in 1632), being in 19$^{\mathrm{th}}$ one of leading international universities in the Northern Europe. When the university was re-established (1802) the professorship of statistics was created at once, and for some periods it seemed to be one of leading disciplines in the faculty. If in Universities of Moscow and St Petersburg the deep theoretical problems of probability were solved, then Tartu was concerned to solving more practical tasks of statistics.

When the first session of International Statistical Institute took place (1883), then among about 50 most well-known statisticians of world there were two men who have worked in the University of Tartu. One of them was Ernst Louis Étienne Laspeyres (1834–1913), one of creators of index-theory in economics who worked in Tartu as professor (1869–1873).

The second was Wilhelm Hector Richard Lebrecht Lexis (1837–1914), well known as the author of Lexis diagram the bivariate plot explaining and helping recalculations of events from individual age to absolute time and *vice versa*. This diagram lies in all elementary books of demography also today.

About the same time the in the University of Tartu a special group of Biostatics (in German: *Biostatiks*) worked. They used statistical methods for solving practical problems – demography, hygiene etc. The most remarkable result of their work is a series of dissertations on demography defended 1860–1886. These dissertations have been made using the same methodology on population Tartu, Viljandi, Narva, Põltsamaa etc.

The fact that the University of Tartu has been a center of biostatistics has been proved also by the fact, that the dissertation of P. D. Enḱo, containing the first epidemiological model, was defended in Tartu, as in leading universities of Russia the importance of the task was not understood.

Also today, when very different registers exist and very special data are collected from samples and expanded to populations, the censuses are organized and the census data are quoted as the most valuable data for all applications. To get comparable data from almost the whole world the list of variables and methodology of data-collection are very seriously regulated internationally.

If we compare the first censuses carried throw in Baltic area (1881) and the forthcoming census (2011), then we see that, surprisingly, the task has changed much more difficult today.

- If in past the preparatory work took about half year, so now it takes about ten times more.
- If in past most of the population were enumerated during one day, so now we plan for the length of the census period about 6 weeks.
- In 1941 the first data of census were published after 4 months. Now Eurostat has fixed the time 28 months.

Also the ideology of censuses has somewhat changed, as the population and its behavior has changed markedly during the last decades. The things exceptional in past are usual today and *vice versa*. It is very difficult to get a clear snapshot of population today, as many dimensions are fuzzy: it is often difficult to define clearly place of usual residence, the composition of household and family; also such a traditionally fixed variable as sex can change from census to census. It seems new paradigms for census will be necessary.

Dual System Estimation of Drug Abuse

Juha M. Alho

Dept. of Computer Science and Statistics, University of Joensuu, Finland

Users of heavy drugs, such as opiates and amphetamines, have an incentive not to reveal their status in surveys, because of the criminality of the use. Administrative records provide an alternate source of information. In Finland, hospital discharge register and register of criminal offenses provide independent, partial lists of the users. Data from double registration can be combined using capture-recapture, or dual system estimation techniques, to estimate the size of the user population and some of its characteristics. However, it is known that capture probabilities depend on the user's age, sex and possibly other factors. This may lead to the so-called correlation bias. To the extent such background characteristics are observable, their effect can be modeled using binary regression, based on conditional likelihoods. The estimates can be used in a Horvitz-Thompson estimator of population size.

The sampling distribution of the Horvitz-Thompson estimator is skew when the data are sparse, however. This suggests that delta-method based assessment of uncertainty is unreliable. We introduce a Bayesian formulation of the problem. The choice of reference priors for the data generating mechanism and population density are discussed. A Poisson model with expectations from a moving average gamma process is used for the latter. Under this model, unconditional likelihoods of the capture data are practicable. They provide a simple basis for Gibbs sampling. The practical implementation is facilitated by data augmentation. The methods are illustrated with data from the province of Southern Finland, in 2002. Potential uses of the techniques in other demographic contexts are noted.

De Novo Detection of DNA Regulatory Binding Regions Using a Bayesian Random Frame Model

Timo Koski[1], **Jukka Corander**[2],
[1] Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden
[2] Department of Mathematics, Åbo Akademi University, Finland

Identification of regulatory binding motifs within DNA sequences is a commonly occurring problem in computational bioinformatics. A wide variety of statistical approaches have been proposed in the literature to either scan for previously known motif types or to attempt *de novo* identification of a fixed number of putative motifs. Most approaches assume the existence of reliable bio database information to build probabilistic a priori description of the motif classes. No method has been previously proposed for making simultaneous inference about the number of putative de novo motif types and their positions within a set of DNA sequences. As the number of sequenced genomes from a wide variety of organisms is constantly increasing, there is a clear need for such methods. Here we introduce a Bayesian unsupervised approach for this purpose by utilizing recent advances in the theory of predictive classification and Markov chain Monte Carlo computation. Our modeling framework enables formal statistical inference in a large-scale sequence screening and we illustrate it by a set of challenging examples.

Learning from the results of genome-wide association studies

Daniel Gudbjartsson

Decode Genetics, Reykjavik, Iceland

The number of common sequence variants in the human genome that associate consistently and verifyably with disease risk has exploded in the last three years (a catalog available at `http://www.genome.gov/26525384`). This increase in replicable findings, from a handful to hundreds, has been due to the creation of genotyping platforms capable of affordable parallel measurement of several hundred thousand single nucleotide polymorphisms (SNPs) covering most of the human genome. An overview of the current findings will be given which highlights some of the features and shortcomings of the available technology and gives some insights into what can be expected from future genetic studies.

Large scale genotyping efforts and collaborative consortia mean sample sizes are reaching tens or even hundreds of thousands of individuals. For some traits, e.g. adult height, this is leading to an overwhelming number of discoveries and the challenge becomes to present results in an informative manner and to link findings to other sources of information, such as gene annotation.

The accumulation of genotype and phenotype information allows for the evaluation of the effect of new findings on a wide range of phenotypes. Specific examples will be given of how the association of SNPs associating with disease can be better understood by analysing their correlation with biological traits, and how SNPs associating with biological traits have been used as candidates for SNPs associating with related human diseases.

Keywords: Human genetics, genome-wide association tests, disease gene mapping.

I3: Causal inference in genetics

SEMIPARAMETRIC TESTS FOR SUFFICIENT CAUSE INTERACTIONS

Stijn Vansteelandt[1], **Tyler VanderWeele**[2] **and James Robins**[3]
[1] Ghent University, Belgium
[2] University of Chicago, U.S.A.
[3] Harvard University, U.S.A.

Developments in genetic epidemiology have recently stimulated a large interest in tests for gene-gene or gene-environment interaction. Ultimately, geneticists are interested in mechanistic interdependencies between genes or genes and environmental exposures. Such mechanistic, biological or sufficient cause interactions would signal the presence of individuals for whom the disease would occur if both genes (or both gene and environmental exposure) were ‘present’, but not if only one of the two were present. Empirical conditions for sufficient cause interactions are based on the sign of contrasts between conditional disease probabilities, given confounders. Logistic regression models are not satisfactory for evaluating such contrasts because they require separate tests for each level of the confounders and because their inherent non-additivity can induce an inflation of the Type I error. While these problems can be avoided by using Bernoulli regression models with linear link, linear models for dichotomous outcomes are prone to misspecification, to which interaction tests are known to be sensitive. To accommodate this, we develop semi-parametric tests for sufficient cause interactions both under conditional and marginal structural models which postulate the probability contrast of interest in terms of a finite-dimensional parameter, but which are otherwise unspecified. Because estimation is often not feasible in these models due to the curse of dimensionality, we develop ‘multiply robust tests’ under a union model that assumes at least one of several working submodels holds. In the special case of a family-based genetic study in which the joint exposure distribution is known by Mendelian inheritance, the resulting procedure leads to asymptotically distribution-free tests of the null hypothesis of no sufficient cause interaction. The results will be illustrated in an application to detect mechanistic gene-environment interactions.

Keywords: Gene-environment interaction; Gene-gene interaction; Semiparametric inference.; Sufficient causes.

References:

Vansteelandt, S., VanderWeele, T., Tchetgen, E.J. and Robins, J.M. (2008). Multiply robust inference for statistical interactions. J. Am. Statist. Assoc. 103, 1693–1704.

CAUSAL INFERENCE IN GENETIC EPIDEMIOLOGY

Carlo Berzuini

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, UK

In the first part of the talk, we shall discuss some aspects of causal inference via graphical models which we consider relevant to the analysis of complex studies in genetic epidemiology. Particular emphasis will be on problems of identifiability of conditional intervention distributions and dynamic plans. In the second part of the talk, the available technology will be illustrated with the aid of a combined prospective-retrospective study on the genetics of susceptibility to coronary artery disease. Particular attention will be paid to issues of model construction in the light of the available biological knowledge.

I4: Life course epidemiology

Comparing methods for life-course analysis – an example: pathways to adulthood

Mervi Eerola[1], **Noona Kiuru**[2], **Katariina Salmela-Aro**[2]
[1] Methodology Centre for Human Sciences, University of Jyväskylä,
[2] Dept. of Psychology, University of Jyväskylä, Finland

Survival or event history methods have traditionally been used in life-course analysis to consider transitions between relatively few well-defined states. The growing complexity of life-course patterns has raised the question of whether the diversity of patterns is preserved when limiting the set of possible trajectories in advance. Furthermore, multiple demographic processes, such as entry into labor force, educational achievements, family formation and parenthood, are all acting in parallel, and flexible methods to describe and analyze interdependent life course trajectories are needed.

In sociology, there has been considerable interest to apply data mining methods to analyze life-course data. These methods were developed in the 1980's to analyze DNA sequences in large databases. The unit of analysis is then not an event as in EHA but the whole sequence of states. Typically, optimal matching or other alignment methods are used to calculate a distance matrix as a measure of similarity between pairs of sequences. Based on the distance matrix, some clustering algorithm is then used to form a typology of sequence patterns. Finally, explanatory covariates can be used to evaluate differences between the clusters.

In this presentation, we discuss the problems and potentials of sequence analysis from a life-course perspective and compare it with more traditional methods. We illustrate its use in the analysis of pathways to adulthood in a cohort data of university students.

Keywords: Sequence analysis, Event-history analysis, Interdependent trajectories.

References:

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *J. R. Statist. Soc. A, 170.* (pp. 167-183).

Eerola, M. (1994): *Probabilistic Causality in Longitudinal Studies.* Lecture Notes in Statistics *92*. Berlin: Springer-Verlag.

Fiocco, M., Putter, H. and van Houwelingen, H. (2008). Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statist. Med.*. Published online.

Case-cohort and nested case-control studies: Differences and similarities

Sven Ove Samuelsen

Department of Mathematics, University of Oslo, Norway
National Institute of Public Health, Oslo, Norway

Case-cohort and nested case-control are both case-control designs within a time-frame. In case-cohort a sub-cohort is sampled at the outset of the study and cases are identified at event times. In nested case-control studies the controls are sampled from the risk sets of the cases. Traditionally one has put "equality signs" between the design and the method of analysis. For nested case-control analysis is performed formally as matched case-control whereas for case-cohort studies a "pseudo-likelihood" approach has been applied. Comparisons of the designs have generally been carried out under these standard methods.

In this talk I will review alternative methods for case-cohort and nested case-control studies. A basic idea is that sampling of controls is performed for both designs and thus methods from survey sampling such

as inverse probability weighting (Horvitz-Thompson estimation) based on probabilities of being included in the study can be useful. For case-cohort studies this amounts to a minor change in the analysis that is carried out. However, based on background information available on the complete population the sampling fractions can be calculated in several ways. Carefully chosen sampling fractions may improve efficiency. One key-word here is post-stratification, but more advanced methods are being developed. For nested case-control studies such weighting methods can in addition help overcome some deficiencies. For instance in a competing risk situation controls sampled to one type of case (disease I) can not be used as controls for cases of another type of case (disease II) with the standard type of analysis. The standard analysis is also linked to the chosen time-scale. By the use of weighting techniques such restrictions are removed. However, as it is common in nested case-control studies to match on several variables in addition to being in the risk set, such as year of birth and area of residence, sampling fractions may sometimes be difficult to obtain.

Size at birth, adult body mass index, and risk of myocardial infarction in the SHEEP study

Kristiina Rajaleid[1,2], **Imre Janszky**[2], **Johan Hallqvist**[2,3]
[1] Centre for Health Equity Studies, Stockholm University/Karolinska Institutet, Sweden
[2] Department of Public Health Sciences, Karolinska Institutet, Sweden
[3] Department of Public Health and Caring Sciences, Uppsala University, Sweden

Background. Small size at birth, an indicator of poor foetal nutrition, is associated with increased risk of cardiovascular disease (CVD) later in life. According to the thrifty phenotype hypothesis foetal undernutrition ŞprogramsŤ the bodyŠs organ structure and function in order to maintain long-term metabolic thriftiness. Later life exposure to overnutrition then results in changes in insulin metabolism and action and increased susceptibility to CVD.
Aim. To test the hypothesis that the effect of impaired foetal growth on CVD in men and women is dependent on body size in adulthood.
Methods. Based on Stockholm Heart EpidEmiology Program (SHEEP) – a population based case-control study of risk factors for incident myocardial infarction (MI); 1058 cases and 1478 controls were included in the current analysis. Unconditional logistic regression modelling was used. The association of standardised birth weight for gestational age with incident MI was explored. Secondly, the interaction effect of standardised birth weight for gestational age with adult body mass index (BMI) at three measurement times separately was assessed. Subsequently, all BMI measures were combined to a life-time BMI trajectory, assuming linear growth/decrease of BMI between the measurement points. Latent class growth analysis, allowing for a quadratic growth factor and with sex as a covariate was used to identify homogeneous trajectory classes within the population. Using the posterior probabilities of the model, each person was allocated to their most likely trajectory class. The interaction analysis with standardized birth weight for gestational age was then repeated using the life-time BMI trajectory class instead of the dichotomised single BMI measurements as the measurement of adult body size. Synergy index was calculated for the interaction effects.
Results. Having a very low standardized birth weight for gestational age was associated with increased risk of incident MI in men and women. Having small size at birth in combination with high BMI at the age of 20, at the time of the MI or as the life-time maximum value of BMI was associated with extremely high risk of MI. We could not establish if it is being overweight at some specific age or a high life-time BMI that in combination with impaired foetal growth predicts MI. Confounding and/or mediation by socioeconomic circumstances at birth and in adulthood, adult CVD risk factors, markers of metabolic syndrome and increased CVD risk in the family explained only a minor part of the increased risk due to interaction.

Spatial Statistics in Epidemiology. Cluster of Disease: Making or Solving Problems

Geir Aamodt

Norwegian Institute of Public Health, Oslo, Norway.

Detection of geographic clusters is an important part of epidemiology. Significant increases of incidences in limited areas are of both public and personal concern and there are needs for efficient statistical methods to help detecting high-risk areas; spatial clusters or space-time clusters.

In this talk, I will first discuss etiological factors of chronic diseases like auto-immune diseases and how clustering might help finding some new clues for these diseases. Typically, the diseases show a significant temporal increase during the last 10-20 years but also some spatial clustering. Many etiological factors have been proposed to explain such changes, and is seems as if microbial variability could be one explanation for these clusters.

Next, I will present different statistical methods for cluster detection and the methods' performances. We will look at global tests, which are developed for detecting but not identify the cluster, such as nearest neighbour tests. We will also look into methods aiming at identifying the clusters (local clustering tests) such as spatial scan methods, nonparametric regression models, and Bayesian disease mapping. We will see that the global tests for spatial randomness are powerful, but identifying the geographic extension of the cluster is more problematic. As cluster change from circular to more elongate, they are more difficult to find. Likewise, the risk difference between the high-risk area and the normal risk area is also important. The statistical methods are also more appropriate in some situations than others.

We will also look at space-time extensions of these methods and evaluate their performance. These methods are not as good evaluated as the spatial cluster detection methods.

Different evaluations show that identifying clusters of disease incidences and their geographic extensions is difficult. Disease clusters are still hard to find.

Keywords: geographic clusters, chronic diseases.

Advances in Gaussian processes for spatial epidemiology

Aki Vehtari

Helsinki University of Technology, Finland

Gaussian processes (GP) can be used to define elaborate spatial and spatio-temporal priors using combination of different covariance functions. Also inclusion of explanatory variables with non-linearities and implicit interactions is straightforward. The challenges with using Gaussian process models are the computational burden and memory requirements. In case of non-Gaussian observation model other difficulty is the analytically intractable inference. In this talk, I will discuss the benefits of the Gaussian processes for modeling latent spatial phenomena. In addition, I will discuss properties of different approximations which can be used to reduce the computation and memory requirements and show how it is possible to achieve both globally and locally good approximations for covariance functions and efficient inference with non-Gaussian observation models. To illustrate the flexibility of the developed methods I will briefly describe two cases. The first case is spatial variation of alcohol-related diseases in 10500 inhabited 5km×5km grid cells in Finland, which is also used to illustrate the modeling of two spatial phenomena having different lengthscales. The second case is spatio-temporal variation of cancer incidences in Finnish counties over 50 years having 20000 data points, which also illustrates the prediction of the future.

III Contributed talks

C1: Bioinformatics I

BAYESIAN CLUSTERING AND FEATURE SELECTION FOR CANCER TISSUE SAMPLES

Pekka Marttinen[1], **Samuel Myllykangas**[2] **and Jukka Corander**[3]
[1] Helsinki University of Technology, Finland
[2] University of Helsinki, Finland
[3] Abo Akademi University, Finland

The versatility of DNA copy number amplifications for profiling and categorization of various tissue samples has been widely acknowledged in the biomedical literature. For instance, this type of measurement techniques provides possibilities for exploring sets of cancerous tissues to identify novel subtypes. The previously utilized statistical approaches to various kinds of analyses include traditional algorithmic techniques for clustering and dimension reduction, such as independent and principal component analyses, hierarchical clustering, as well as model-based clustering using maximum likelihood estimation for latent class models.

While purely algorithmic methods are usually easily applicable, their suboptimal performance and limitations in making formal inference have been thoroughly discussed in the statistical literature. Here we introduce a Bayesian model-based approach to simultaneous identification of underlying tissue groups and the informative amplifications. The model-based approach provides the possibility of using formal inference to determine the number of groups from the data, in contrast to the ad hoc methods often exploited for similar purposes. The model also automatically recognizes the chromosomal areas that are relevant for the clustering. Validatory analyses of simulated data and a large database of DNA copy number amplifications in human neoplasms are used to illustrate the potential of our approach. Our software implementation BASTA for performing Bayesian statistical tissue profiling is freely available for academic purposes at `http://web.abo.fi/fak/mnf/mate/jc/software/basta.html`.

Keywords: Cancer biology, DNA amplifications, unsupervised classification and feature selection.

References:

Marttinen, P., Myllykangas, S. and Corander, J. (2009). Bayesian clustering and feature selection for cancer tissue samples. *BMC Bioinformatics 10:90.*

RECONSTRUCTING POPULATION HISTORIES FROM SINGLE-NUCLEOTIDE POLYMORPHISM DATA

Jukka Sirén[1], **Jukka Corander**[2] **and Pekka Marttinen**[3]
[1] Department of Mathematics and Statistics, University of Helsinki, Finland.
[2] Department of Mathematics, Åbo Akademi University, Finland.
[3] Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, Finland.

Population genetics encompasses a strong theoretical and applied research tradition on the multiple demographic processes that shape genetic variation present within a species. When several distinct populations exist in the current generation, it is often natural to consider the pattern of their divergence from a single ancestral population in terms of a binary tree structure. Inference about such population histories from molecular data has been an intensive research topic in the most recent years. The most common approach uses coalescent theory (Hein *et al.*, 2005) to model genealogies of individuals sampled from the current populations. Such methods are able to compare several different evolutionary scenarios and to estimate demographic parameters. However, their major limitation is the computational complexity associated with the indirect modelling of the demographies.

In this work we propose a novel fully Bayesian method for inferring population histories from unlinked single-nucleotide polymorphisms. We use an approximation to the neutral Wright-Fisher diffusion to model random fluctuations in allele frequencies (Nicholson *et al.*, 2002). The population histories are modelled as binary trees that represent the historical order of divergence of the different populations. A combination of analytical, numerical and Monte Carlo integration techniques are utilized for the inferences.

Keywords: population history, single-nucleotide polymorphism, genetic drift

References:

Hein J., M.H. Schierup MH, C. Wiuf (2005). *Gene Genealogies, Variation and Evolution.* Oxford: Oxford University Press.

Nicholson, G., A.V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, P. Donnelly (2002). Assessing population differentiation and isolation from single nucleotide polymorphism data. *J. R. Stat. Soc. B 64*, 695–715.

MEASURING GOODNESS OF SEGMENTATION PERFORMED WITH HIDDEN MARKOV MODELS

Jüri Lember[1] **and Alexey Koloydenko**[2]
[1] University of Tartu, Estonia
[2] Royal Holloway, University of London, United Kingdom

Hidden Markov models are commonly used in computational molecular biology. The classical examples involve modeling DNA-sequences or proteins. Often, the main object of interest is the true realization of the hidden Markov chain. Estimation of the hidden realization is also known as segmentation. For example, DNA sequences can be segmented into coding and non-coding regions, or into CpG islands and background. A widely used estimate of the underlying hidden path is maximum-likelihood, or maximum a posteriori (MAP), sequence of states, commonly known as Viterbi alignment. An alternative is pointwise (symbol-by-symbol) maximum-likelihood, or PMAP-alignment. This latter alignment is optimal in the sense of minimizing the expected number of incorrectly classified elements of the hidden sequence. When the HMM extends to infinity, then (under certain mild conditions) both, Viterbi and PMAP alignments converge, generally to distinct limiting alignments. Viewed as proper random processes derived from the original hidden Markov process, the infinite limiting alignments can be used to study various asymptotic properties of their respective segmentations. In particular, we show how this asymptotic analysis can be used for measuring goodness of and improving segmentation.

Keywords: Hidden Markov models, Viterbi alignment, regenerative processes, semi-supervised learning, active learning.

A Simple Stochastic Cell Cycle Model with Application to Cytometry Data

Rimantas Eidukevičius[1] **and Dainius Characiejus**[2] **and Ramūnas Janavičius**[3]
[1] Faculty of Mathematics and Informatics, Vilnius University, Lithuania
[2] Institute of Immunology, Vilnius University, Lithuania
[3] Department of Human and Medical Genetics, Vilnius University, Lithuania

There are various methods to analyze cell kinetics based on deterministic models.

A simple method to estimate means and variances of durations of cell cycle phases G_1 (preparation for the DNA synthesis), S (DNA synthesis) and G_2M (preparation for mitosis and division into two daughter cells) is proposed. A semi-Markov process with three states is used to model the cell cycle, i.e. the sojourn times are independent with arbitrary distributions, for example, exponential, shifted exponential, etc. A special case is continuous time Markov chain.

This semi-Markov model is fitted to flow cytometry data obtained after BrdUrd (5-bromo-2'-deoxyuridine) pulse labelling of cells in the synthesis phase. Propidium iodide (PI) is used to estimate total amount of DNA. The cellular DNA content and the amount of incorporated BrdUrd are measured simultaneously using flow cytometry.

The method of estimation of cell cycle stage durations is based on flow cytometric analysis of one or two tumour samples after BrdUrd labelling. A minimal number of samples taken at different times after the BrdUrd labelling is equal to the number of parameters of the transition distributions.

Because of small number of the available samples only a simple stochastic process is used to model the cell cycle. The rest phase G_0 is accessible only from the state G_1, proliferating cells are moving from G_1 to S, from S to G_2M. After mitosis the fraction $a \in [0, 2]$ of daughter cells enter G_1 and other cells enter G_0. An assumption $a = 1$ is used for simplicity. It corresponds to so called "rectangular" age distribution of proliferating cells. This assumption is substantiated by the horizontal shape of the PI histogram in the S stage.

C2: Spatio-temporal modeling

ASSESSING THE UNCERTAINTY OF THE POLYGONAL DECLUSTERING ESTIMATOR OF A SPATIAL MEAN

Juha Heikkinen

Finnish Forest Research Institute, Vantaa, Finland

In polygonal declustering, a spatial average is estimated by a weighted mean with weights proportional to the areas of polygons in a Voronoi tessellation generated by the locations of the sample points (e.g., Isaaks & Srivastava 1989, ch. $\sim$ 10). Sampling error is usually assessed either through a variogram model fitted to the sample or by using the formula related to simple random sampling (SRS). In the former approach, the effects of model mis-specification or poorly estimated variogram are difficult to predict, while the latter can easily yield overly pessimistic estimates of precision.

In the case of systematic sampling, variance estimates for the sample mean can be derived from balanced differences between values observed at neighbouring sample points (see Cochran 1977, sec. 8.11, for a brief review and further references). Typically, the variance is still over-estimated, but the upward bias is often much smaller than with the SRS formula.

The aim of this work was to generalize the method of balanced differences to the case of arbitrarily distributed sample points, essentially providing a model-free alternative to the variogram-based error estimators for polygonal declustering. The research was motivated by the need to generalize the sampling error estimator of the National Forest Inventory of Finland from the square grid design into more general ones. Performance of the polygonal declustering estimator and the potential for estimating its variance is evaluated through a simulation study imitating realistic inventory conditions.

An issue of specific interest is the performance in the presence of strong trends over the inventory area. Then, due to edge effects, unweighted mean may not be the optimal estimator of the spatial average even in the case of systematic sampling (see Cochran 1977, sec. 8.6). An edge corrected estimator suggested by Yates (1948) is an analogue of polygonal declustering in one dimension.

Keywords: Geostatistics, Systematic sampling, Variance estimation, Forest inventory.

References:

Cochran, W.C. (1977). *Sampling Techniques.* New York: John Wiley & Sons.

Isaaks, E.H., R.M. Srivastava (1989). *An introduction to Applied Geostatistics.* Oxford University Press.

Yates, F. (1948). Systematic Sampling *Phil. Trans. Roy. Soc. London A 241*, 345–377.

SPATIAL-TEMPORAL MODELING OF BALTIC SEA COASTAL ZONE PARAMETERS

Laura Saltyte, Kestutis Ducinskas

Department of Statistics, Klaipeda University, Lithuania

A modeling technique named spatial connection, which can be applied to univariate time series data observed at different spatial locations (spatial time series) is proposed. Spatial connection is implemented by two methods: 1st method is based on averaging of the parameters of ARIMA models with spatial weights; 2nd - based on formal application of an ordinary kriging procedure to the model parameters. The modeling techniques are applied to the hydrological and hydrochemical parameters in Baltic Sea coastal zone received from Lithuanian Sea research center.

Keywords: spatial connection, spatial time series, spatial weights, ARIMA, kriging, semivariogram

IMPROVED ESTIMATION IN A SPATIO-TEMPORAL GROWTH-INTERACTION MODEL

Ottmar Cronie

Chalmers University of Technology & University of Gothenburg, Sweden

The spatio-temporal growth-interaction model has recently been studied in (Särkkä and Renshaw, 2006). The basis of the process is an immigration-death process governing the number of arrivals of individuals (occuring according to a Poisson process) who's locations are uniformly distributed in the region of interest. This process also governs the individuals' deaths (individuals are assigned iid exponential life-times). Upon arrival an individual is assigned a mark which changes with time according to a deterministic growth equation which consists of two parts, an individual growth term and a spatial interaction term. The growth and interaction parameters are estimated using the method of least squares. The immigration and death parameters have thusfar been estimated separately using Maximum likelihood (ML) estimators.

We here consider a less approximative way of estimating the immigration parameter and the death parameter. By exploiting the Markov property of the process one can find the transition probabilities and thereby the full likelihood. Since no closed form solutions of the ML estimators are available, numerical maximization is used to to find the estimates. If sufficient time is available a brief overview of the edge correction methods employed in the least squares estimation will be given.

Keywords: Spatio-temporal marked point process, Immigration-death process, Maximum likelihood estimation.

References:

Särkkä, A., Renshaw, E. (2006). The analysis of marked point patterns evolving through space and time. *Computational Statistics & Data Analysis 51*, 1698–1718.

EUROPEAN BIODIVERSITY IS STILL IMPACTED BY NATURE - SPATIAL AUTOCORRELATION IN MIXED MODELS

Jaan Liira[1], **Josef Settele**[2] **and Martin Zobel**[1]

[1] Institute of Botany and Ecology, Tartu University, Estonia

[2]UFZ - Centre for Environmental Research Leipzig-Halle, Department of Community Ecology,Germany

The impact of anthropogenic drivers on biodiversity in Europe has frequently been discussed. We performed an analysis to elucidate the integrated effects of a comprehensive set of anthropogenic and natural drivers on the large-scale biodiversity of some of the best-known species groups, and to give suggestions for the most effective biodiversity indicators. We studied the relationship between taxon richness and anthropogenic drivers or geographical location, applying the general linear mixed model analysis with spatial-autocorrelation settings. Second, we searched for the most effective large-scale indicators of biodiversity of various taxonomic groups. We show that different taxa have different spatial range of aut-correlation, that both natural and anthropogenic drivers are significant determinants of biodiversity - latitudinal and longitudinal gradients of biodiversity predominated, but the effect of habitat loss became evident as well. Apart from birds, we showed that species richness within one taxonomic group is a more useful indicator for other groups than solely anthropogenic and natural drivers.

C3: Causal inference

Bayesian nonparametric monotonic regression

Olli Saarela[1] **and Elja Arjas**[2,1]
[1] National Institute for Health and Welfare, Helsinki, Finland
[2] Department of Mathematics and Statistics, University of Helsinki

Bayesian approaches to nonparametric modeling of monotonic regression functions, either in the univariate case or in the generalised additive model (GAM) framework, have been presented by, e.g., Arjas and Gasbarra (1994), Neelon and Dunson (2004), Dunson (2005) and Brezger and Steiner (2008). We present a generalisation which allows Bayesian estimation of monotonic regression surfaces of an arbitrary shape over one or more covariates. Our monotonic construction is based on a marked point process formulation, where the random point locations and the associated marks (function levels) together define a piecewise constant realisation of the random regression surface. Statistical inferences are then based on the posterior samples of these random functions. As in the traditional monotonic regression problem, we define monotonicity using a partial ordering of the values of the regression function. A clear advantage of our Bayesian approach, when combined with Markov chain Monte Carlo methods, is that we only need to consider the partial ordering constraints locally. Our method can accommodate any type of likelihood, and can be attached as a part of a larger probability model. It also allows reduction into lower dimensional submodels, then acting as a device in model selection.

Keywords: Monotonic regression, nonparametric Bayesian modeling

References:

Arjas, E., D. Gasbarra (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica 4*, 505–524.

Brezger, A., W.J. Steiner (2008). Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. *Journal of Business & Economic Statistics 26*, 90–104.

Dunson, D.B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association 100*, 618–627.

Neelon, B., D.B. Dunson (2004). Bayesian isotonic regression and trend analysis. *Biometrics 60*, 398–406.

Predictive Bayesian inference and dynamic treatment regimes: the MACS data revisited

Elja Arjas[1,2] **and Olli Saarela**[2]
[1] Department of Mathematics and Statistics, University of Helsinki, Finland
[2] National Institute for Health and Welfare, Helsinki, Finland

Dynamic treatment regime is a decision rule in which the choice of a treatment at any given time can depend on the known past history of an individual, including baseline covariates, earlier treatments, and their measured responses. In this talk it is argued that finding an optimal regime can, at least in moderately simple cases, be accomplished by a straightforward application of nonparametric Bayesian modeling and predictive inference. As an illustration we consider the well-known Multicenter AIDS Cohort Study (MACS) data set (see, for example, Moodie et al. 2007), studying the effect of AZT initiation on CD4 cell counts during a 12-month follow-up.

Keywords: Dynamic treatment regime, nonparametric Bayesian modeling, predictive inference.

References:

Moodie, E.E.M., T.S. Richardson, D.A. Stephens (2007). Demystifying optimal dynamic treatment regimes. *Biometrics 63*, 447–455.

Major treatment incompliance (including unplanned "cross-over" of study treatment): structural nested failure time model as an alternative analysis method?

Peter Jakobs[1] **and Pasi Korhonen**[2]
[1] Novartis Oncology, Basel, Switzerland
[2] University of Turku, Turku and StatFinn Oy, Espoo, Finland

The basics of structural nested failure time models (SNFTM) for survival endpoints will be introduced. The SNFTMs attempt to estimate the treatment effects while accounting for post-randomisation changes in the study treatments. Post-randomisation treatment changes may include for instance dose-adjustments, discontinuation of study treatment or crossing-over to other randomized treatments. If there exists time-dependent confounders which are risk factors for the clinical outcome and affected by past study treatment and at the same time further predict the future treatment decisions, the conventional methods for adjusting such confounders are biased. The SNFTMs allow unbiased estimation the underlying treatment effect under certain assumptions. The concept will be illustrated by a randomized placebo-controlled phase III study for *Exjade* (an approved Novartis oral iron chelator) in patients with myelodysplatic syndromes (MDS) and transfusional iron overload.

Keywords: Structural nested failure time models, intermediate confounding variables, counterfactuals

A principal stratification-based model to estimate the effect of blinding in a clinical trial with open and blind arms.

Krista Fischer

MRC Biostatistics Unit, Cambridge, United Kingdom

In the Estonian Postmenopausal Hormone Therapy (HT) trial, the participating women were randomized to two blind arms (HT or placebo) and two non-blind arms (HT or untreated control). Before coming to the recruitment visit, the randomized women were informed on whether they were randomized to blind or non-blind sub-trial. About 50% withdrew at this stage, whereas withdrawal was more likely on blind arms.

The final analysis was conducted using the data on recruited women only, showing significant differences in the main outcomes (incidence of cancer, bone fractures, cerebrovascular and cardiovascular diseases) between blind and non-blind arms.

Such effects of blinding can have two possible reasons: either they arise from differential pre-recruitment selection of women to blind and non-blind arms or there are some post-recruitment effects (possibly explained by the effects of blinding on women's and physician's behavior).

Assuming that each randomized woman belongs to one of the three principal strata depending on her willingness to participate in a blind or non-blind trial, the two types of blinding effects are defined as between- and within-strata differences. To estimate them, a modeling approach will be proposed, where identifiability is achieved using baseline covariates that predict selection.

Application of these ideas on the data demonstrates that (depending on outcome) there is some evidence that both, pre- and post-recruitment effects of blinding are present. These results have important implications on generalizability of results from clinical trials, especially when conducted on healthy volunteers.

Keywords: Causal Inference, Randomized Clinical Trial, Principal Stratification

C4: Statistics in epidemiology

Feature extraction in visualizing and describing a predictive classifier: A case study

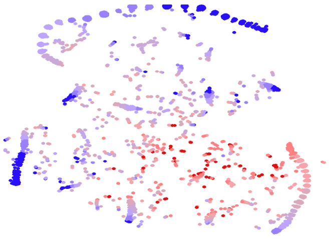
Elina Parviainen,[1] **Aki Vehtari**[1]
[1] Dep. of Biomedical Engineering and Computational Science, Helsinki University of Technology, Finland

In this case study, we work on a classification of patient data. We have trained a predictive classifier and want to study the explanations of the obtained results in more detail. We need to decide which covariates to emphasize when describing the results and we want to see if some groups of patients should be described separately. With a nonlinear model and high-dimensional data these decisions can be difficult to make. This kind of analysis is especially relevant for health care data, which describes the attributes of human beings. Such data is inherently complex since the same observed behavior can have several possible explanations. For example, discharge from a hospital may in general be predicted well by age, but in groups with an additional disease recovery may be slower.

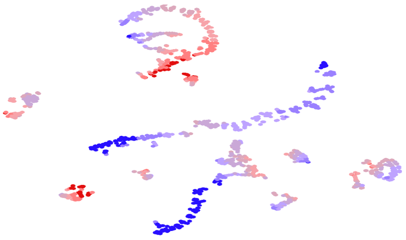
Visualization using dimension reduction to 2D or 3D space aids in the analysis, but only if the low-dimensional presentation shows some relevant structure. If a dimension reduction method tries to preserve structure in all dimensions equally faithfully, the visualization may end up being hard to interpret. We want the visualization to emphasize those covariates contributing to prediction. For achieving this, we exploit the features already learned by a classifier network. When we do dimension reduction in the feature space instead of using raw data, also the visualization can show structure relevant to classification.

Keywords: feature extraction, visualization, dimension reduction

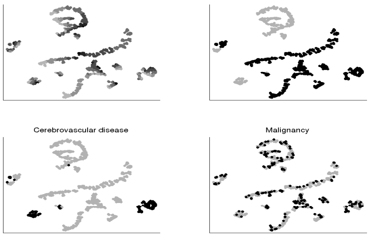
As dimension reduction on the original data does not reveal any clear structure, ...



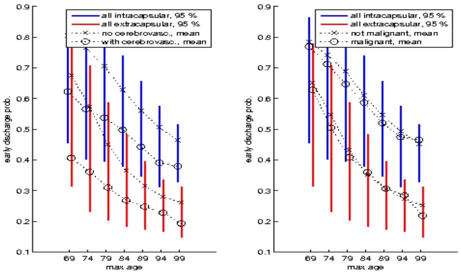
... we would rather use the result of performing dimension reduction on features extracted from a predictive classifier model.



As the clusters tend to have similar values for good predictors and more variation for others, we can study the distibutions of covariates in clusters...



... to quickly recognize factors having higher or lower impact on class probability.



Modelling and prediction of weekly influenza A specimens in England and Wales

Jurate Saltyte Benth

University of Oslo, Norway

We propose a rather simple model, which fits well the weekly human influenza incidence data from England and Wales. A standard way to analyze seasonally varying time-series is to decompose them into different components. The residuals obtained after eliminating these components often do not reveal time dependency and are normally distributed. We suggest that conclusions should not be drawn only on the basis of residuals and that one should consider the analysis of squared residuals. We show that squared residuals can reveal the presence of the remaining seasonal variation, which is not exhibited by the analysis of residuals, and that the modelling of such seasonal variations undoubtedly improves model fit.

Boosting Additive Quantile Regression for Investigating Childhood Malnutrition

Nora Fenske[1] **and Thomas Kneib**[2] **and Torsten Hothorn**[1]
[1] Ludwig-Maximilians-Universität München, Germany
[2] Carl-von-Ossietzky-Universität Oldenburg, Germany

Ordinary linear and generalized linear regression models relate the mean of a response variable to a linear combination of covariate effects and, as a consequence, focus on average properties of the response. When analyzing childhood malnutrition in developing or transition countries based on such a regression model, this implies that the estimated effects describe the mean nutritional status. To focus in more detail on the malnutrition aspect, it is more relevant to investigate the lower quantiles of the response distribution depending on associated risk factors.

In the presented work, we analyze risk factors for childhood malnutrition based on data collected in the 2005/2006 India Demographic and Health Survey. To handle this task, we suggest a semiparametric extension of quantile regression models where nonlinear effects are included in the model equation. For estimation of the resulting additive quantile regression model, we develop, evaluate and apply a novel boosting approach combining quantile regression, as treated in Koenker (2005), with boosting for additive models, as described in Kneib et al. (2009). Our proposal allows for data-driven determination of the amount of smoothness required for the nonlinear effects and provides an automatic variable selection property. The results of our empirical evaluation suggest that boosting is a reasonable tool for estimation in linear and additive quantile regression models and helps to identify yet unknown risk factors for childhood malnutrition.

Keywords: Functional gradient boosting, penalized splines, additive models, variable selection, model choice.

References:

Fenske, N., T. Kneib, T. Hothorn (2009). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Technical Report, No. 52, Department of Statistics, LMU München.

Kneib, T., T. Hothorn, G. Tutz (2009). Variable selection and model choice in geoadditive regression models. Biometrics. To appear (available early online).

Koenker, R. (2005). Quantile regression. Economic Society Monographs. Cambridge University Press, New York.

Factor Analysis of the Metabolic Syndrome Identifies two Factors with Different Survival Patterns in Elderly

John Öhrvik[1], **Pär Hedberg**[2], **Tommy Jonason**[2], **Ingemar Lönnberg**[3], **Göran Nilsson**[34]
[1]Department of Medicine, Karolinska Institutet, Stockholm
[2]Department of Physiology, Central Hospital, Västerås
[3]Department of Medicine, Central Hospital, Västerås
[4]Department of Clinical Research, Uppsala University, Sweden

Background: Factor analysis reduces a set of directly measured variables into a smaller set of underlying factors representing unique statistically independent domains termed 'factors'. Previous factor analyses of the continuous components of the Metabolic Syndrome (MetS) have identified 2-4 factors in middle aged people. Many of these reports blood pressure as one factor.

Design and Methods: We performed a factor analysis of the continuous components of MetS in a cohort of 198 men and 203 women comprising 65% of all 75-year-olds from the city of Västerås in Sweden. Varimax rotation, which results in high loadings for a few components and near zero loadings for the rest, was applied to get a more easily interpretable solution. To study the accuracy of the factor loadings 10-fold cross-validation was applied. The cut-off for the factor loadings was set to 0.4. Prospective associations of these factors with all cause mortality (median follow-up 10.6 person-years) were assessed by multiple Cox proportional hazard regression. The best subset approach, using Akaike information criterion (AIC) as performance measure, was used to find an optimal set of significant confounders. The predictive ability of the original components and the derived factors was assessed by the time dependent area under the ROC curve (AUCt).

Results: The MetS consistently comprised two factors applying 10-fold cross-validation. Factor 1; the metabolic factor, consisted of waist, HDL-cholesterol, triglycerides and fasting glucose. Factor 2; the blood pressure factor, consisted of systolic and diastolic blood pressure. These two factors explained in median: 57.9%; range: 56.9-58.5% (1st factor 31.2%; 30.0-32.0% and 2nd factor 26.6%; 26.1-27.1%) of the total variation in men. The corresponding figures for women were 63.0%; 60.8-64.7% (1st factor 36.2%; 34.0-37.6% and 2nd factor 27.0%; 26.7-27.2%). The factor loadings showed consistent patterns over all the 10 different splits. Only in one of the male cases the factor loading for fasting glucose was slightly below the cut-off (0.39<0.40). During 10-years follow-up 91 (46%) men and 58 (29%) women died. The metabolic factor was significantly related to 10-years mortality in men; Hazard Ratio (HR) 1.30 (95%CI 1.07-1.57, p=0.008) and nearly significantly in women; HR 1.27 (95%CI 0.99-1.63, p=0.058). In a pooled analysis the HR for the metabolic factor was 1.19 (95% CI:1.02-1.39, p=0.027) adjusting for sex, previous myocardial infarction, known hypertension and current smoking, the only significant confounders using AIC in a best subset approach. The AUCt increased from 0.583 to 0.688 adding the metabolic factor to the confounders. The blood pressure factor was not significantly related to survival either crude or adjusted.

Conclusion: Factor analysis of the basic variables of the MetS among 75-year-olds from a general population identified a metabolic and a blood pressure factor. The former factor was significantly related to 10-years survival and the relation remained after adjusting for sex, previous myocardial infarction, known hypertension and current smoking.

C5: Statistics in ecology and forestry

ESTIMATION OF DISCHARGE RATING CURVES WITH B-SPLINES

Kristinn Mar Ingimarsson[1,2],**Birgir Hrafnkelsson**[1], **Sigurdur M. Gardarsson**[1] **and Arni Snorrason**[2]
[1] School of Engineering and Natural Sciences, University of Iceland, Iceland
[2] Icelandic Meteorological Office, Iceland

Rivers across the globe are monitored for flood watch and to quantify their potential energy for the generation of hydroelectricity. For these purposes water level measurements are collected over time with automatic water level recorders. However, discharge is usually of more interest than water level but the cost of water level measurements is small relative to direct discharge measurements. Discharge rating curves are used to transform water level to discharge and are one of the fundamental tools in hydrology. To estimate discharge rating curves, paired measurements of these two variables are needed.
The standard power-law relationship is given by $q = a(w - c)^b$ where q is the discharge, w is the water level, a is a scaling parameter, b is a shape parameter and c is equal to the water level giving zero discharge and the relationship is not defined for $w < c$. The standard power-law relationship is sufficient to describe the relationship between discharge and water level for most rivers. However, for some rivers the relationship deviates from the standard power-law. In this case, the common approach is to use multi-segment rating curve which assumes the standard power-law in each segment and continuity at the endpoints of the segments.
We propose a method based on the Bayesian approach and cubic B-splines to model discharge data which deviate from the standard power-law. This method is different from using a multi-segment rating curve. The parameters a and b are modeled as smooth functions of water level with cubic B-splines and with constants. A lognormal model is assumed for the discharge measurement. The parameter c is an unknown constant. The model for the logarithm of the discharge is a linear model in the coefficients of the cubic B-splines. The priors for the unknown constants in the models for a and b are based on knowledge about the strong negative relationship between a and b found in previous data. Markov random field priors are selected for the coefficients of the two cubic B-spline functions. The linear form and careful selection of the priors lead to an efficient and stable Markov chain Monte Carlo for the generation of samples from the posterior. One of the advantages of the proposed model is its parameterization which is such that it can reduce to the standard power-law if needed. The model is tested on several data sets from the Icelandic Meteorological Office.

Keywords: Bayesian estimation, Markov random fields, the standard power-law for discharge, water level.

A TEMPORAL COMPOSITIONAL ANALYSIS OF WATER QUALITY MONITORING DATA

Mark J Brewer[1], **Dörthe Tetzlaff**[2], **Iain A Malcolm**[3] **and Chris Soulsby**[2]
[1] Biomathematics and Statistics Scotland, Aberdeen, United Kingdom
[2] School of Geosciences, University of Aberdeen, United Kingdom
[3] FRS Freshwater Laboratory, Perthshire, United Kingdom

An important part of understanding the behaviour of river catchments is determining the geographical sources of water flow and how the variation in relative proportions contributed by different sources changes over time and in response to drivers. The method of end-member mixing analysis (EMMA) has been developed in the hydrological literature (e.g. Genereux, 1998), often relying on solving mass-balance equations for which assessments of uncertainty are poorly defined. Recent work in the statistical literature has been concerned with Bayesian analysis of compositions (e.g. Brewer et al., 2005), and we discuss extensions of that work which model the source distributions directly in the absence of suitable end-member samples.

Our motivating data set was collected for two streams at Loch Ard, central Scotland, over an 18 year period from 1988. Data were recorded approximately weekly until 1996 and fortnightly thereafter, in the form of measurements of alkalinity as tracer and of the rate of flow for both streams. Previous work (e.g. Tetzlaff et al., 2007 for this catchment) has generated source distributions simply by using the corresponding alkalinity values for the highest and lowest flows, but there are problems with this method: the sample sizes can be very low and as a result, identification of end-member distributions is often unreliable. Our method using Weibull mixed models for source identification has proven to be more robust.

Keywords: Compositional analysis; Water quality; Time series.

References:

Brewer, M.J., Filipe, J.A.N., Elston, D.A., Dawson, L.A., Mayes, R.W., Soulsby, C. and S.M. Dunn (2005). A hierarchical model for compositional data analysis. *Journal of Agricultural, Biological and Environmental Statistics*, **10**, 19–34.

Genereux, D. (1998). Quantifying uncertainty in tracer-based hydrograph separations. *Water Resources Research*, **34**, 915–919.

Tetzlaff, D., Malcolm, I.A. and C. Soulsby (2007). Influence of forestry, environmental change and climatic variability on the hydrology, hydrochemistry and residence times of upland catchments. *Journal of Hydrology*, **346**, 93–111.

TEST OF THE FAMILY OF RTH-FUNCTIONS FOR GROWTH AND DISTRIBUTION MODELS

Artur Nilson

Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Tartu, Estonia.

The tests of different members of the family of asymptotic rth-functions

$$rth(x,c,r,s,t,u,v) = \left\{1 - \left[\frac{r}{((\exp(x/c))^s - 1)^t} + r\right]^u\right\}^v$$

with six possible shape parameters r, c, s, t, u and v proved its fitness for modeling forest growth and stem diameter distribution. Reasonable number on variable parameters was four or three with the rest of parameters fixed to one or the choice for special subset of data series. Different four-parameter variants occurred in the top of the best four parameter growth functions for modeling 690 stand height and 664 volume growth series from forest yield tables. The three best with marginal difference were the functions by A. Levaković (1935), D. Todorović (1961) together with one of the rth-functions. Variants with three variable parameters with the fourth fixed as fitted for the subset of data by tree species were the best among the three parameter ones. Many of well-known growth functions (Weber, Weibull, Mitscherlich, Todorović, Schubert, Bertalanffy etc with parameter $r = 1$ or $r = 2$ for Sevastjanov) can be described as special cases of the rth-function thus helping to enhance coherence in the system of growth models. The set of user-defined functions in the R and MSVisualFoxPro© environment for computing the derivative, inverse function and inflexion point proved handy for estimating the increment or pdf, quantiles and mode or growth culmination point correspondingly. The tests for modeling tree diameter distribution gave similar results.

Keywords: rth-functions, growth function, distribution function, forest, modeling

TEASING OUT THE WORKINGS OF COMMUNITY DYNAMICS

Crispin M. Mutshinda[1] **and Robert B. O'Hara**[1]
[1] Department of Mathematics and Statistics, University of Helsinki, Finland

Understand the interplay of environmental stress, food web interactions and demographic stochasticity in driving community structure and dynamics is a fundamental goal in ecology. A sensible approach to this issue is to design mechanistic models capable of teasing out the underlying factors. We use a hierarchical Bayesian approach to develop such a model, which further encompasses a stochastic search variable selection methodology for identifying the set of biologically sensible interactions by automatically shrinking the coefficients of irrelevant interactions toward zero. We illustrate the model implementation with long-term macro-moth (*Lepidoptera*) light-trapping data from the Rothamsted Insect Survey network in the UK.

Keywords: community drivers, environmental forcing, food web interaction, stochastic search variable selection

Statistical detection of cut-off cadmium concentration in breast cancer etiopathogenesis study

Rima Kregzdyte[1,2], **Loreta Strumylaite**[1], **Dale Baranauskiene**[1], **Oleg Abdrkhmanov**[1]
[1] Institute for Biomedical Research, Kaunas University of Medicine, Lithuania
[2] Department of Preventive medicine, Kaunas University of Medicine, Lithuania

Background. Breast cancer is the most common cancer in women worldwide and in Lithuania as well. Many studies are performed in order to find etiopathogenesis of breast cancer. A hospital based case-control study is carried out by researchers of Kaunas University of Medicine. One of its aims is to assess a relationship between cadmium exposure and the disease. The first findings show that cadmium concentration in breast tumor and normal tissue differs statistically significantly.
Objective. The objective of the followed analysis was to determine cadmium concentration in breast tissue over which the cancer is more likely.
Material and methods. The concentration of cadmium was determined in breast tissue samples of 57 breast cancer and 50 benign tumor patients. Tumor and normal tissue close to tumor samples were taken for the analysis. The metal concentration was determined by atomic absorption spectrometry (PerkinElmer, Zeeman 3030). Receiver operating characteristic (ROC) curves were used to detect a cut-off value.
Results. The area under the curve for cadmium concentration in tumor was 0.684 (95% confidence interval 0.579—0.788; $p = 0.001$), that indicated cadmium to be a significant predictor of the disease status. The cut-off point for optimal test was chosen to optimize the rate of true positives and minimizing the rate of false positives. The minimum sum of false positives and false negatives squared was at the concentration of 28 ng/g. Calculated sensitivity was 0.72 and the specificity was 0.66 at this point.
Conclusions. Revised estimate of cadmium cut-off concentration made on enlarged sample size is valuable predictor of breast cancer and is useful in the further studies.

Statistical modelling of measurement errors in gas chromatographic analyses of blood alcohol content

Jukka Corander[1], **Rossana Moroni**[12], **Paul Blomstedt**[1], **Tapani Reinikainen**[2], **Erkki Sippola**[2]
[1] Department of Mathematics, Åbo Akademi University, Finland
[2] National Bureau of Investigation, Vantaa, Finland

Introduction. Over consumption of alcohol is one of the most frequent cause of fatal accidents on the roads according to the United NationsŠ Statistics of road traffic accidents in Europe and North America. In many countries the crime of Driving Under Influence (DUI) has two limits of severity and the offenderŠs penalty is derived by his/her Blood Alcohol Content (BAC). Headspace gas chromatographic measurements of ethanol content in blood specimens from suspect drunk drivers are routinely carried out in forensics laboratories. In the widely established standard statistical framework, measurement errors in such data are represented by Gaussian distributions for the population of blood specimens at any given level of ethanol content. It is known that the variance of measurement errors increases as a function of the level of ethanol content and the standard statistical approach addresses this issue by replacing the unknown population variances by estimates derived from large samples using a linear regression model. Appropriate statistical analysis of the systematic and random components in the measurement errors is necessary in order to guarantee legally sound security corrections reported to the police authority.
Aim. Here we address this issue by developing a novel statistical approach that takes into account any potential nonlinearity in the relationship between the level of ethanol content and the variability of measurement errors.
Methods. Our method is based on standard nonparametric kernel techniques for density estimation using a large database of laboratory measurements for blood specimens. Furthermore, we address also the issue of systematic errors in the measurement process by a statistical model that incorporates the sign of the error term in the security correction calculations.

Results. Analysis of a set of spiked-in blood samples demonstrates the importance of explicitly handling the direction of the systematic errors in establishing the statistical uncertainty about the true level of ethanol content. Use of our statistical framework to aid quality control in the laboratory is also discussed.
Conclusion. Using the density estimation technique allows us to take into account the fact that the SD may not be a constant in the population of blood samples (as assumed in Jones and Shuberth, where the SD was provided by the regression model for any mean BAC). In details, our SD distribution represents the uncertainty about the SD in the population; as a consequence, the level of confidence is valid and conservative also for those individuals whose samples are subject to a higher level of variation.

PRACTICAL ASPECTS OF ASSESSING AGREEMENT OF CLINICAL MEASUREMENT METHODS

Bendix Carstensen[1]
[1] Steno Diabetes Center, Niels Steensens Vej 2-4, DK-2820 Gentofte, Denmark, & Department of Biostatistics, University of Copenhagen,
bxc@steno.dk, www.biostat.ku.dk/~bxc

The comparison of two methods of measurement using the so-called ”Bland-Altman” procedure of plotting the difference against the mean for each pair of observations has become the *de facto* standard for analysis of method comparison studies without replicates [1,2].
This scenario is expanded to comparing two or more methods of measurement with arbitrary replication structure [3], linking methods by linear functions. For measurements by method m on individual i and replicate r we use:

$$y_{mir} = \alpha_m + \beta_m\big(\mu_i + a_{ir} + c_{mi}\big) + e_{mir},$$
$$a_{ir} \sim \mathcal{N}(0,\omega^2), \quad c_{mi} \sim \mathcal{N}(0,\tau_m^2), \quad e_{mir} \sim \mathcal{N}(0,\sigma_m^2)$$

This is a non-linear random effects model. But it is not easily implemented in standard software. An iterative fitting algorithm is suggested, but it is actually more feasible to recast it in a graphical models framework using the BUGS machinery. The BUGS routines allows the user to fit variance component models, generate translation formulae between methods with proper prediction limits accounting for all sources of variation, and is not restricted to comparing only two methods.
I will give some illustrative examples and demonstrate the R-package `MethComp` that provides fitting algorithmns for the models and BUGS routines, and provides a friendly and flexible interface to estimation and reporting of models for method comparison studies.

Keywords: Method comparison, conversion formulae, calibration, reproducibility, variance components

References:

JM Bland and DG Altman (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310.

JM Bland and DG Altman (1999) Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, **8**, 136–160.

B Carstensen, J Simpson and LC Gurrin (2008) Statistical models for assessing agreement in method comparison studies with replicate measurements *International Journal of Biostatistics*, **4**(1), article 16.

Comparing Experimental Designs for Benchmark Dose Calculations for Continuous Endpoints

Kristi Kuljus[1], **Dietrich von Rosen**[2],
Salomon Sand[3] **and Katarina Victorin**[3]
[1] Swedish University of Agricultural Sciences, Umeå, Sweden
[2] Swedish University of Agricultural Sciences, Uppsala, Sweden
[3] Karolinska Institutet, Stockholm, Sweden

The BMD (benchmark dose) method which is used in risk assessment of chemical compounds is based on dose-response modelling. To take uncertainty in the data and model fitting into account, the lower confidence bound of the BMD estimate (BMDL) is suggested to be used as a point of departure in health risk assessments. The main aim of the present study was to investigate whether an increased number of dose groups and at the same time a decreased number of animals in each dose group improves conditions for estimating the benchmark dose.

We have used Bayesian design approach for comparing experimental designs for benchmark dose calculations in the case of continuous endpoints. The approach is exemplified by considering the class of Hill models. Since Hill models are nonlinear, the optimum design for estimating the benchmark dose depends on the values of the unknown parameters. For this reason we have considered Bayesian designs and assume that the parameter vector has a prior distribution. A natural design criterion is to minimize the expected variance of the BMD estimator.

We present an example where we have calculated the value of the design criterion for several designs and try to find out how the number of dose groups, the number of animals in the dose groups and the choice of doses affects the criterion value for different Hill curves. It follows from our calculations that to avoid the risk of unfavorable dose placements, it is good to use designs with more than four dose groups. We can also conclude that any additional information about the expected dose-response curve, e.g. information obtained from studies made in the past, should be taken into account when planning a study, because it can improve the design.

Keywords: benchmark dose, dose-response modelling, Hill models, optimum designs.

References:

Kuljus, K., D. von Rosen, S. Sand, K. Victorin (2006). Comparing Experimental Designs for Benchmark Dose Calculations for Continuous Endpoints.
Risk Analysis 26, 1031–1043.

C7: Bioinformatics II

LOCAL REVERSE-COMPLEMENT SYMMETRY OF DNA SEQUENCES

Marijus Radavičius[1] **and Jurgita Židanavičiūtė**[2]
[1] Institute of Mathematics and Informatics, Vilnius, Lithuania
[2] Vilnius Gediminas Technical University, Vilnius, Lithuania

DNA is a duplex structure comprising two self-complementary strands which are held together via hydrogen bonds between nucleotides in a base pair. There are two base pairs: a nucleotide A on one strand always binds with T on the other and, respectively, C binds with G. The simplest hypothesis of DNA strand similarity states that frequencies of nucleotides of the same base pair are approximately equal within single DNA strands. Since the second strand is read in the reverse order an extension of this first-order similarity to higher-orders is called reverse-complement symmetry or simply strand symmetry (Baisnée et al., 2002, Zhang and Huang, 2008). Results of empirical studies using asymmetry measures and graphical tools show that for long DNA sequences (approximate) strand symmetry generally holds with rather rare exceptions. Moreover, it is established that a lower-order parity does not imply a higher-order strand symmetry (Baisnée et al., 2002). DNA sequences (segments of DNA strand) can be treated as finite-state random sequences. Simons et al. (2005) have proposed a *global* probabilistic model of strand symmetry and investigated its goodness-of-fit empirically.
In our presentation a *local* model is described. It is assumed that DNA sequence is generated by a Markov random field invariant under the corresponding to strand symmetry group of reverse-complement transformations. By making use of this invariance a convenient parametric form of the partition function of the Markov random field is derived. The proposed model was applied to non-coding regions of bacteria genoms from database *GenBank*. Preliminary results of testing goodness-of-fit confirm empirical observations that long DNA sequences tend to be more symmetric than the shorter ones.

Keywords: DNA, Strand symmetry, Markov random field, Invariance, Generalized logit, Goodness-of-fit.

References:

Baisnée, P.-F., S. Hampson, P. Baldi (2002). Why are complementary DNA strands symmetric? *Bioinformatics. 18*, no. 8, 1021–1033.

Simons, G., Y.-C. Yao, G. Morton (2005). Global Markov models for eukaryote nucleotide data. *Journal of Statistical Planning and Inference 130*, 251–275.

Zhang, S.-H., Y.-Z. Huang (2008). Characteristics of oligonucleotide frequencies across genomes: Conservation versus variation, strand symmetry, and evolutionary implications *Nature Precedings* hdl:10101/npre.2008.2146.1.

A METHOD OF VISUALIZATION OF DNA SEQUENCES

Tomas Rekašius

Vilnius Gediminas Technical University

Genome signature was introduced by Jeffrey (1990). It identifies DNA sequence $W = (w_1, w_2, \dots w_n), w_i \in \mathcal{A} := \{A,C,G,T\}$ with a certain sequence of points $(x_i, y_i), i = 1, \dots, n,$ in a unit square and is very convenient way of visualization and comparison of DNA sequences. However, due to the fractal character of DNA signatures the differences between them are difficult to interpret.
Here the fractal character means that close points in the unit square do not necessarily correspond to similar DNA sequences and vice versa. This effect can by illustrate by following example with $\mathcal{A} := \{0, 1\}$. Let symbolic sequences $W_1 = 1000000000$, $W_2 = 0111111111$, $W_3 = 0000000001$ and $W_4 = 1111111110$ be sequences of coefficients in dyadic representations of $x_1 = 1/2, x_2 = 1/2 - 2^{-10}, x_3 = 2^{-10}$ and $x_4 = 1 - 2^{-9}$, respectively. Then the both pairs (W_1, W_2) and (W_3, W_4) are very different however $|x_1 - x_2| = 2^{-10}$ whereas $|x_3 - x_4| > 1 - 2^{-8}$.

Let d be a distance function between two symbolic (DNA) sequences. Distance between unit square points is a Euclidean. Our goal is to construct a "continuous" mapping of DNA sequences to the unit square. This problem can be formulated as a multidimensional scaling problem (Borg, Groenen, 2005). The distance d between symbolic sequences is usually defined as edit (Levenshtein) distance (Waterman, 1995). It is not very suitable distance measure for DNA sequences because edit operations (partially) ignore a complex structure of interaction of adjacent symbols. A new measure of similarity of symbolic sequences taking into account their complexity is proposed. Some applications to visualization of bacterial genomes from GenBank are presented and discussed.

Keywords: Genome signature, distance measure, visualisation of nucleotide sequences.

References:

Borg, I., Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications (Second Edition)*. Springer.

Jeffrey, H.J. (1990). Chaos Game Representation of Gene Structure. *Nucleic Acids Res.* Vol. 18, (pp. 2163-2170).

Waterman M.S. (1995). *Introduction to Computational Biology.* Chapman & Hall, London.

ELUCIDATING PREDICTABLY DIFFERENT PHENOTYPES BY MULTICLASS CLASSIFICATION AND CLUSTERING

Erinija Pranckeviciene, Justas Arasimavicius, Valentina Gineviciene and Vaidutis Kucinskas[1]

[1] Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Lithunia

Reliable associations between the genotype and phenotype enable to account for a potential of a development of certain traits. Once the markers, associated with quantitative or binary traits, are determined, they can be used for testing/prediction. In sports genetics, several variants are hypothesized being associated with the general athlete performance and endurance phenotypes: *ACE, PGC1A, ACTN3* and *PPARA*. An elucidation of predictively different phenotypes, related to the specific combinations of the alleles in these variants, would advance the training strategies and recommendations in the early sportsmen carrier. We present an early preliminary report on a computational methodology for finding a mapping between the predictably different phenotypes and genotypic variants. The experiments were carried out using 21 standard continuous phenotypic features of 193 Lithuanian athletes, typed for *ACE, PGC1A, ACTN3* and *PPARA* variants. The phenotypes of the athletes are different in various sports categories: speed and strength, mix, endurance and team. By linear discriminant analysis the phenotypes are classified into four different categories, in which the phenotypes separate linearly. The four typed markers in total produce 3^4 unique genotypes. In our dataset we observed only 48. A distribution of the genotypes within the classified phenotype categories allows an assignment of a degree of a genotype membership in every category. Hence the individual genotype is characterized by four membership values. Hierarchical clustering of these values produces a mapping of the unique genotypes into the predictably distinct phenotype categories.

Keywords: multiclass classification, clustering, genotype-phenotype association, *ACE, PGC1A, ACTN3, PPARA*, sports genetics.

References:

Gineviciene, V., V. Kucinskas, J. Kasnauskiene (2009). The angiotensin converting enzyme gene insertion/deletion polymorphism in Lithuanian professional athletes *Acta Medica Lithuanica.* In press.

Williams, A.G., J.P. Folland (2008). Similarity of polygenic profiles limits the potential for elite human physical performance *J Physiol. 586.1*, 113–121.

Ziegler, A., I.R. Konig, J.R. Thompson (2008). Biostatistical aspects of genome-wide association studies *Biometrical Journal. 50*, 8–28.

C8: Statistical modeling

WHERE ARE THE COWS AND WHAT ARE THEY DOING

Søren Højsgaard[1]
[1] Aarhus University, Denmark

This work is motivated by the fact that moderne dairy production is very intensive in many countries with farms with several hundred cows and a minimal number of staff to run such farms. Furthermore, there is an increased focus on animal welfare among the general public. In this connection automatic monitoring of the cows is essential. Four main focus areas are: 1) Detection of oestrus, 2) detection of eating disorders 3) detection of lameness and 4) detection of mastitis.

Moderne sensor technology is becomming cheaper and cheaper and it is therefore feasible to monitor individual cows online by attaching various sensors either to the cows or to feeding boxes and milking machines.

In an ongoing project we have 1) recordings of positions of cows determined by a bluetooth based positioning system, 2) recordings of accelerations of cows measured in three dimensions, 3) measurements of their lying/standing behaviour and other activities and 4) inline measurements of certain substances in milk.

State space models and other types of latent variable time series models are natural candidates for modelling such data. In the talk we will focus on one or two specific models applied to these problems.

Keywords: dairy cows, exponential smoothing, hidden semi–Markov models, online monitoring, state space models

EFFECT OF A PEN IN GROUP FEEDING TRIALS: MODELING DATA FROM SUCKLER COWS

L. Jauhiainen[1], M. Manninen[2], J. Üfversten[1]
[1] MTT Agrifood Research Finland, Services Unit, Jokioinen, Finland
[2] MTT Agrifood Research Finland, Animal Production Research, Jokioinen, Finland

The current study provides information on the extent and containment of intraclass correlation concerning group feeding trials on suckler cows. The research comprised the re-analysis of six previously reported experiments and an adherent simulation study. Intraclass correlation coefficients were estimated for eight variables. They were seen to be higher in experiments on bulls compared with those on cows and more notable in the measurements of live weight gain compared with final weight. Moreover, the intraclass correlation coefficients were generally high for all variables measuring behavioural patterns. The simulation study showed that using a single animal as an experimental unit could be seen as valid in certain situations, but it could not be extended to cover all cases. The simulation study also showed that the common mixed model approach had significant problems when the intraclass correlation was slight. Degrees of freedom adjustment methods should be used when the mixed models are fitted to the data. In general, the research strengthened the arguments that much more effort should be placed on the planning and statistical analysis of group feeding experiments, especially in behavioural studies.

Keywords: intraclass correlation, mixed model, degrees of freedom

THE PERFORMANCE OF BEST LINEAR UNBIASED PREDICTION IN SMALL RANDOMISED COMPLETE BLOCK EXPERIMENTS

Johannes Forkman[1]
[1] Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

In mixed models, the method of Best Linear Unbiased Prediction (BLUP) is known to give smaller mean square error than the Best Linear Unbiased Estimation (BLUE), provided that the variance components are known. In practice, the variance components are estimated, and empirical best linear unbiased prediction (EBLUP) is used instead of BLUP. Predictions using EBLUP should be close to predictions using BLUP if the variance components are sufficiently well estimated. In small experiments, this requirement is not fulfilled. It has been argued that EBLUP is adequate when analysing large data sets, when precise estimation of the relevant variance components is possible, while BLUE is preferable when the information in the data about the variance components is small.

Agricultural field experiments and other biometrical experiments are often designed as randomised blocks. The usual statistical model for analyses of randomised block experiments includes two factors: blocks and treatments. The block factor could be regarded as fixed if the blocks are complete, or otherwise random for recovery of inter block information. It may be more difficult to choose whether to model the treatment factor as fixed or random.

In this study, results when modelling treatments as random in small randomised complete block experiments with normally distributed treatment effects is compared with results obtained when modelling treatments as fixed. We study, by simulation, the root mean square error and the performance of prediction intervals and statistical tests when the numbers of blocks and treatments are small.

Keywords: BLUP, random effects, mixed model, randomised block experiment.

IDEAS ABOUT LIKELIHOOD-BASED DATA ANALYSIS IN R

Peter Dalgaard[1]
[1] Department of Biostatistics, University of Copenhagen

R has for a long time provided access to good quality optimizers. In combination with its powerful programming language, it is often quite straightforward simply to write up a likelihood and find maximum likelihood estimators by direct maximization, provided the problem is “sufficiently well-behaved”. This can be more expedient than trying to fit a problem into an existing modelling framework such as `glm()`. However, there is a need for support functionality at several levels. The `mle()` function in R and various associated functions were written to provide a basic toolkit for summarizing an ML fit, extract quantities, and describing the behaviour of the likelihood near the optimum (profiling).

It is desirable to extend the toolkit in various ways. For instance, the `mle()` function has no checks that its `mll` argument is actually a negative log-likelihood, and it is easy to make mistakes. Also, it would be nice to have a way to construct models from simpler “building blocks”.

This suggests that we need an extended likelihood-model class, objects from which the `mll` function can be extracted when needed, but which share some aspects with the more conventional model classes, and can be defined using simpler primitives than low-level programming. Some natural operations may be defined on models (e.g., combination of two experiments, mixture models, or censoring), and it is necessary to discuss the requirements this puts on the object structure.

Keywords: R, likelihood, model-building

C9: Statistical genetics

Bayesian QTL mapping based on reconstruction of recent genetic histories

Dario Gasbarra[1], Matti Pirinen[1], Mikko J Sillanpää[1] and Elja Arjas[1]
[1] Department of Mathematics and Statistics,
University of Helsinki,
Finland

We describe a Bayesian model that simultaneously incorporates the recent genealogical history and a quantitative phenotype measurement of the sampled individuals. The genealogical history consists of the pedigree and the gene flow at certain marker loci (Gasbarra et al. 2007). The phenotypic variance is modeled as a linear combination of contributions from (a random number of) affecting loci, the polygene effect and the residual effect. This work extends the framework of the variance component linkage analysis to the settings where the pedigrees need not be known beforehand, but where they are included in the model as latent variables.

As an input the method requires unphased genotype data at the marker loci, some estimates for the population allele frequencies and for the marker map, as well as some parameters related to the population size and the mating behaviour. Given such data the posterior distribution of the phenotype parameters (the number, the locations and the relative effects of the affecting loci) is studied by using the reversible jump Markov chain Monte Carlo methodology. We also introduce two shortcuts related to the phenotype parameters that let us do analytic integration, instead of stochastic sampling, at some parts of the algorithm.

The method is tested on two simulated data sets, which clearly show that the ability to model the unobserved recent history is advantageous in gene mapping. Comparisons with traditional variance component linkage analysis and association analysis are also carried out and good results are obtained.

Keywords: linkage analysis, association analysis, pedigree estimation, IBD estimation

References:

Gasbarra, D., Pirinen, M., Sillanpää, M.J., Arjas, E. (2007). Estimating Genealogies from Linked Marker Data: A Bayesian Approach. *BMC Bioinformatics 8:411.*

Estimating Heritability of Fluctuating Asymmetry in Sticklebacks

Robert B. O'Hara

Department of Mathematics and Statistics, University of Helsinki, Finland

Fluctuating asymmetry (random left/right asymmetry) has had a checkered history in its use as a measure of individual fitness. In part this is because of the difficulties in its estimation, as large sample sizes are needed, and observation error has to be controlled. Here I use a large data set from sticklebacks to estimate FA and its heritability.

Sticklebacks are armoured with lateral plates, which help protect them from predators. There is variation in the number of plates they carry, and this can demonstrate FA. A large crossing experiment was carried out, with a half-sib design. Over two thousand fish were scored for the presence of plates at each of 30 myomeres (in essence each position on the body) on each side of the fish.

In order to estimate FA, i.e. the propensity for a fish to have a plate on only one side of a myomere, the effect of the number of plates had to be removed. Hence, a bivariate model, with the overall odds of platedness and the random asymmetric variation were both modelled with a genetic model. In addition, the effect of eda, an (unobserved) major gene, had to be inferred. A hierarchical Bayesian model was constructed and fitted to the data. This is not straightforward: it requires two genetic models, plus the direction of any asymmetry has to be estimated individually for each myomere.

We found moderate heritability for the number of plates and for FA in platedness, as well as the expected variation in platedness along the fish, and a large effect of the eda locus. The flexibility of the hierarchical model allowed us to extract the effects of the different factors affecting platedness of sticklebacks, including the focal effect of FA.

COMPARISON OF BAYESIAN MODELS FOR GENOMIC SELECTION USING REAL DAIRY DATA

K.L. Verbyla[1,2,3], **P. Bowman**[1], **B. Hayes**[1], **H. Raadsma**[4], **M. Khatkar**[4], **and M.E. Goddard**[1,2,3]
[1] Department of Primary Industries, Victoria, Australia
[2] The University of Melbourne, Victoria, Australia
[3] The Cooperative Research Centre for Beef Genetic Technologies, Armidale, NSW, Australia
[4] Cooperative Research Centre for Innovative Dairy Products, Centre for Advanced Technologies in Animal Genetics and Reproduction, University of Sydney, NSW, Australia

The information provided by dense marker maps and an increase in genotyping has allowed the development of multiple QTL models that allow mapping of quantitative traits. These multiple QTL models can be utilised as a selection technique by simultaneously evaluating the marker effects across the genome. The effects of these markers are summed to produce an animal's genome-assisted breading value (GEBV), and this process is termed genomic selection. Genomic selection assumes the markers are in linkage disequilibrium (LD) with the QTL. This study evaluated seven Bayesian approaches for predicting SNP effects for genomic selection by assessing their ability to accurately predict GEBV in a real dairy data set containing 1498 Australian Holstein-Friesian bulls genotyped for 39048 SNPs. The methods included BLUP (infinitesimal model assumptions) and a Bayesian approach in which each SNP had a SNP specific variance. Each had three variants using a selected subset of SNPs (weighted and unweighted) and all SNPs. The other approach, Bayes C used Stochastic Search Variable Selection. The different approaches were applied to estimate GEBV for 9 traits. Validation populations containing bulls proven in each of 2005, 2006 and 2007 were used to assess the accuracy of the GEBV by comparing the estimated GEBV with the published Australian Breeding Values. The results showed that all models produced accuracies that were not significantly different for most traits. This suggests that reduced parameterisation could be used to produce equally accurate GEBV. However for the trait fat percentage, there was a significant difference between the methods with BLUP not performing well. This difference can be explained by the true distribution of QTL effects compared to the prior distributions for the SNP effects and the SNP variances for each method. BLUP does not appear to perform well for those traits that have a known gene or genes responsible for explaining a large percentage of genetic variation for the trait. This suggests that any prior information about a trait's QTL effect distribution should be used to determine which model will produce the most accurate GEBV.

Keywords: Bayesian analysis, genomic selection, prediction

Mean and median survival times of cancer patients should account for informative censoring and general mortality predictions

Karri Seppä[1,2] **and Timo Hakulinen**[1]
[1] Finnish Cancer Registry, Helsinki, Finland
[2] Department of Mathematical Sciences, University of Oulu, Finland

The mean and median of patient survival times are used to summarize the survival experience of a group of patients. When the estimation of the mean and the median survival time is based on incomplete follow-up information, the main sources of bias are informative censoring and use of the most recent general population mortality to predict the asymptotic survival of the patients after the point of cure or the last point of follow-up.

Weighted averages of the age-specific results are used for bias reduction and predicted general mortality tables for improving estimates of the patients' asymptotic mortality. As examples, colon cancer patients diagnosed in 1970–1979 and thyroid cancer patients diagnosed in 1978–1987 were followed until the end of each diagnosis period (incomplete follow-up) and until the end of 2005 (almost complete follow-up).

Due to informative censoring, the crude estimates of the mean lifetime grossly overestimate the survival of the colon cancer patients and underestimate the survival of the thyroid cancer patients. Together with the most recent population life tables, the bias-reducing method succeeds in estimating the mean and the median lifetime accurately.

Stratifying by age is essential when the mean or median lifetime of the patients with a wide age-range is to be estimated. The bias-reducing method should be used if a single summary estimate for the whole patient group is needed. The median is preferable if more than half of the patients die soon after diagnosis. Predicted population life tables should be used in extrapolation.

Keywords: Mean lifetime, Median lifetime, Incomplete follow-up, Informative censoring, Bias, Population-based study.

References:

Seppä, K., Hakulinen, T. (2009). Mean and median survival times of cancer patients should be corrected for informative censoring. *J Clin Epidemiol.* [Epub ahead of print]

VISUALIZING COVARIATES IN PROPORTIONAL HAZARDS MODEL

Juha Karvanen[1]
[1] National Institute for Health and Welfare, Helsinki, Finland

A graphical method called the rank-hazard plot (Karvanen and Harrell, 2009) visualizes the relative importance of covariates in a proportional hazards model. The key idea is to rank the covariate values and plot the relative hazard as a function of ranks scaled to interval $[0, 1]$. The relative hazard is plotted with respect to the reference hazard, which can be e.g. the hazard related to the median of the covariate. Transformation to scaled ranks allows plotting of covariates measured in different units in the same graph, which helps in the interpretation of the epidemiological relevance of the covariates. Rank-hazard plots show the difference of hazards between the extremes of the covariate values present in the data and can be used as a tool to check if the proportional hazards assumption leads to reasonable estimates for individuals with extreme covariate values. Alternative covariate definitions or different transformations applied to covariates can be also compared using rank-hazard plots.

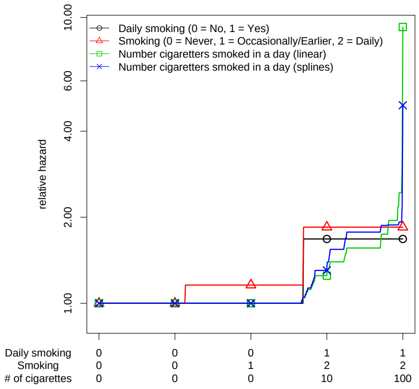


Figure 1. Comparison of proportional hazards models where the risk of coronary heart disease is explained by smoking covariate that has been defined in alternative ways. The rank-hazard curves of the other covariates in the models, non-HDL cholesterol blood pressure and BMI, are not displayed.

Keywords: coronary heart disease, rank-hazard plot, relative importance, statistical graphics, survival analysis, visualization

References:

Karvanen J., Harrell F.E. (2009). Visualizing covariates in proportional hazards model. *Statistics in Medicine*, in press.

ASYMPTOTIC PROPERTIES OF COLLECTIVE CONDITIONAL LIKELIHOOD ESTIMATORS FOR BAYESIAN NETWORK CLASSIFIERS WITH CENSORED DATA

Priyantha Wijayatunga and Xavier de Luna

Department of Statistics, Umea University, Umea, Sweden

Bayesian networks are competitive classifiers, especially when their parameters are estimated with maximizing the collective conditional likelihood (CCL) function based on the densities $p(y \mid x)$, $\forall x$ where Y is the classifying variable and X is the vector of predictors. However the maximization has no closed form solution. The collective conditional likelihood of θ given N number of data cases on (Y, X), say, $\left\{\left(y^{(j)}, x^{(j)}, c^{(j)}\right)\right\}_{j=1}^{N}$ is

$$CCL_N(\theta) = \prod_{j=1}^{N} \left\{p_\theta\left(y^{(j)} \mid x^{(j)}\right)\right\}^{1-c^{(j)}} \left\{P_\theta\left(Y \geq y^{(j)} \mid x^{(j)}\right)\right\}^{c^{(j)}}$$

where $c^{(i)}$ is 1 if y is censored and 0 otherwise for $i = 1, ..., N$. Here we extend the results of Wijayatunga and Mase for the case of censored training data. Note that censored training data are common in survival analysis, econometrics, etc.
We will also discuss briefly how to measure the degree of dependence between random variables and how it can be used for the selection of predictors.

Keywords: Bayesian network classifiers, censored data, collective conditional likelihood estimators, asymptotic properties, degree of dependence.

References:

Wijayatunga, P., S. Mase (forthcoming). Asymptotic properties of maximum collective conditional likelihood estimator for naive Bayes classifiers, *International Journal of Statistics and Systems*

Bootstrap Confidence Intervals for Dynamic Path Models

Szilárd Nemes[1] **and Dragi Anevski**[2]
[1] Division for Clinical Cancer Epidemiology, Sahlgrenska Academy, University of Gothenburg
[2] Centre for Mathematical Sciences, Lund University with Lund Institute of Technology

Path analysis is a recursive causal system, and can be seen as an extension of multivariate regression. Dynamic Path Models are based on counting process framework. Let $T \geq 0$ a positive random variable, δ an indicator variable for the event of interest and $N(t) = 1\{T \geq t, \delta = 1\}$ the corresponding counting process. The additive hazard model can be written as d$N(t)$ = d$B_0(t) + \sum_i$ d$B_i(t)X_i$ + d$M(t)$ where d$B_i(t)$ are regression functions of the additive hazards model and d$M(t)$ is a martingale increment (Andrersen $et.al.$ 1992). Also define $\psi_{hj}(t)$ as a least squares regression coefficient when $X_j(t)$ is regressed onto $X_h(t)$. The cumulative indirect, direct and total effects of a Dynamic Path Model are given by

$$\begin{aligned}
\text{ind}(Z_h(t) \rightarrow N(t)) &= \sum_{i=1}^{r} \int_0^t \left(\prod_{l=1}^{w_i - 1} \psi_{i,i_{l+1}}(s) \right) \text{d}B_{i_{w_i}}(s) \\
\text{dir}(Z_h(t) \rightarrow N(t)) &= B_h(t) \\
\text{tot}(Z_h(t) \rightarrow N(t)) &= B_h(t) + \sum_{i=1}^{r} \int_0^t \left(\prod_{l=1}^{w_i - 1} \psi_{i,i_{l+1}}(s) \right) \text{d}B_{i_{w_i}}(s).
\end{aligned}$$

Path analysis for time dependent data was recently introduced by Fosen $et.al.$ (2006). The large sample distributions of the estimators are unknown and bootstrap confidence intervals are suggested as a tool of assessment. It is however unclear how bootstrap samples should be drawn. Here we present and discuss the applicability of non-parametric, parametric and weird bootstrap strategies alongside with technicalities and implementation issues.

Keywords: counting process, survival analysis, additive hazards model, time-dependent covariates

References:

Andersen P.K., Borgan Ø., Gill R.D. and Keiding N. (1992) Statistical Models based on Counting Processes. Springer-Verlag.

Fosen J.,Ferkingstad E., Borgan Ø. and Aalen O.O. (2006b). Dynamic Path Analysis - A New Approach to Analyzing Time-Dependent Covariates, *Lifetime Data Analysis 12*: 143-167.

Poster presentations

BAYESIAN PREDICTIVE CLASSIFIERS

Yaqiong Cui[1], **Jukka Sirén**[2], **Jukka Corander**[1] **and Timo Koski**[3]
[1]Åbo Akademi University, Finland,[2]University of Helsinki, Finland,
[3]Royal Institute of Technology, Sweden

A general Bayesian classification framework is introduced for sets of items which are characterized in terms of measured values in multiple finite alphabets using predictive representations based on random urn models and generalized exchangeability. The predictive framework allows for more expressive representations of uncertainty than those typically used under the standard latent class formulation, including the handling of missing observations among training or unknown samples. A generalization of an earlier introduced unsupervised classification model based on random urns is here derived for supervised and semi-supervised situations. A marginal predictive classifier based on the maximum *a posteriori* rule for each item separately is shown to be asymptotically equivalent with a joint classifier allocating all items simultaneously when the amount of training data tends to infinity. Choice of a particular classification framework is shown to have consequences for the classification uncertainty assessment. The computational challenges associated with the predictive classification representations and their possible generalizations are also discussed.

Keywords: Bayesian inference, clustering, exchangeability, predictive inference, probabilistic classification

References:

Corander, J., Gyllenberg, M. and Koski, T. (2007). Random Partition models and Exchangeability for Bayesian Identification of Population Structure. *Bull. Math. Biol.* **69**, 797-815.

Geisser, S. (1966). Predictive discrimination. In Krishnajah, P.R. (Ed.). Multivariate analysis. New York and London: Academic Press.

SPATIO-TEMPORAL ANALYSIS OF DISEASE INCIDENCE WITH SPARSE GAUSSIAN PROCESSES

Jouni Hartikainen[1], **Jarno Vanhatalo**[1], **Aki Vehtari**[1] **and Eero Pukkala**[2]
[1] Department of Biomedical Engineering and Computational Science, Helsinki University of Technology, Finland
[2] Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland

Gaussian processes (GPs) are probabilistic models for unknown functions attractive for modeling risk surfaces in the context of disease mapping. The traditional approach is to model the disease occurence (i.e. incidence or mortality) as a Poisson process with spatially varying rate, which is a product of expected occurrence count and relative risk, for which the latter is given a prior structure for smoothing the risk surface across the spatial domain. Here, the GP based spatial priors are extended into a spatio-temporal domain for extracting the temporal dynamics of the spatial risk distribution. Due to flexibility of the model a wide variety of covariance structures can be specified, for instance, separate spatial and temporal components as well as an effect, in which space and time are coupled.

The drawback of the GP based models is the heavy computational burden, which results in cubic and quadratic computational time and memory space requirements, respectively. In the recent years several approaches have been proposed for mitigating these problems by $sparsifying$ the covariance structure in varying ways. In this presentation, we discuss how to construct sparse approximations explicitly in spatio-temporal setting for yielding fast and accurate estimates. When combined with deterministic

approximate inference methods for latent values the method allows computing of results for data sets of over 20000 data points in few hours with a standard desktop PC.
To illustrate the usefulness of the proposed method we present analysis of real cancer data sets containing municipal level disease counts of Finland in years 1953-2007. The analysis shows that by breaking the covariance structure to separate spatial, temporal and spatio-temporal components the fit gets clearly better. We also discuss the forecasting property of the constructed model framework in the light of the given data sets.

Keywords: spatio-temporal models, Gaussian processes, disease mapping.

REPRODUCIBLE STATISTICAL ANALYSIS WITH STATWEAVE

Søren Højsgaard[1] **Russel Lenth**[2]
[1] Department of Genetics and Biotechnology, Aarhus University, Denmark
[2] Department of Statistics and Actuarial Science, The University of Iowa, USA

Literate programming originates in computer science but is a very helpful concept in statistical analyses. The idea behind literate programming is to create software as works of literature: Embed source code into descriptive text (rather than the opposite which is common practice) Thereby the software can follow the flow of thoughts and logic and can be more readable by humans.
When applied to statistical analysis this means one will create a single source document consisting of descriptive text and chunks of program code, for example R, SAS or Stata code. When the source document is processed (this process is called weaving), the code chunks are executed. The results, for example tables and graphics, are then subsequently automatically combined with the descriptive text in a single target document. This makes the analysis completely reproducible and it is straight forward to re-run an analysis if data is modified.
We will present a program called StatWeave which facilitates literate programming with statistical software. The document formats can be either LaTeX or OpenDocument Text (OpenOffice). The supported (statistical) programs include R, SAS, S-plus, Stata, Maple and Matlab. It is possible to use more than one of these in a single source file.

CLIMATIC SIGNALS EXTRACTED FROM RING-WIDTH CHRONOLOGIES OF SCOTS PINE FROM ESTONIA

Maris Hordo

Department of Forest Management, Estonian University of Life Sciences, Tartu

Dendrochronological and dendroclimatological methods are useful tools for revealing the dominant factors influencing radial tree-growth. Dendrochronology refers to the use of tree rings to date events and dendroclimatology as the science of reconstructing past climate by use of tree rings. Tree rings reflect environmental conditions and their changes and store the reaction pattern over time, which can be later used as an archive. In this study I analyze Scots pine (Pinus sylvestris L.) radial growth responses to climatic factors in heath and mesotrophic forest site types in Estonia. For this purpose event years of the radial growth were estimated. In 2007 increment cores from Scots pine living trees were collected from an Estonian network of research plots. Altogether 889 trees from 109 sample plots were cored, but in this study data from 37 heath and 72 mesotrophic plots were used, 287 and 602 trees respectively. We hypothesize that extreme climate conditions – severe winter, cold spring and also summer drought are the main causes of abrupt decrease of a tree radial growth.

Keywords: radial growth, climate, pointer years, Scots pine, tree-ring chronology

Estimating location and variability parameters from classified potato tuber size data

Timo Hurme[1], **Lauri Jauhiainen**[1] **and Jukka Öfversten**[1]
[1] MTT Agrifood Research Finland, Services Unit, Method Services, FI-31600 Jokioinen, Finland

In several applications in the field of agriculture the information for each statistical unit is recorded in a form of frequency distribution, even though the underlying phenomenon is continuous by nature. If the upper and lower limits for each level are defined precisely, estimation of location and variability is straightforward. However, typically in practise the first level consists of all the values smaller than the level's upper limit and similarly the last level consists of all the values larger than the level's lower limit. The problem arises because it is not possible to discover the midpoints of these levels. We propose a simple approach to estimate the unknown location and variation parameters of the underlying distribution function assuming only the general type of the distribution function (e.g. Gaussian). To illustrate the problem we used the Finnish statutory variety trial data from 1973 – 2003 where the tubers were classified into four size groups (under 35 mm, 35 – 55 mm, 55 – 70 mm, over 70 mm). A preliminary graphical inspection implied that the underlying distribution of the size of the tuber was normal. Our results showed that over the years the average size of tuber had increased from 48.0 mm to 51.5 mm. At the same time the standard deviation of the tuber size had decreased from 9.4 mm to 8.8. mm. This change was statistically notable because usually variation tends to increase as the average increases. We also found that there were significant differences between the current potato varieties in tuber size and its standard deviation. This result may turn out to be important for practical farming because the price of the yield usually depends on these characters.

Keywords: tuber size distribution, classified data, location, variability.

References:

Bussan, J.A., P.D. Mitchell, M.E. Copas, M.J. Drilias (2007). Evaluation of the Effect of Density on Potato Yield and Tuber Size Distribution *Crop Science 47*, 2462–2472.

Travis, K.Z. (1987). Use of a simple model to study factors affecting the size distribution of tubers in potato crops *J. agric. Sci., Camb. 109*, 563–571.

Life course analysis on the effects of *FTO* on the adulthood BMI in the Northern Finland Birth Cohort 1966

Marika Kaakinen[1,2], **Esa Läärä**[3], **Anneli Pouta**[4], **Anna-Liisa Hartikainen**[5], **Anja Taanila**[1], **Ulla Sovio**[6], **Mark McCarthy**[7,8] **and Marjo-Riitta Järvelin**[1,2,4,6]
[1] Institute of Health Sciences, University of Oulu, Finland
[2] Biocenter Oulu, University of Oulu, Finland
[3] Dept. of Mathematical Sciences, University of Oulu, Finland
[4] Dept. of Child and Adolescent Health, National Institute for Health and Welfare, Finland
[5] Dept. of Obstetrics and Gynaecology, University of Oulu, Finland
[6] Dept. of Epidemiology and Public Health, Imperial College London, UK
[7] Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, UK
[8] Wellcome Trust Centre for Human Genetics, University of Oxford, UK

Many studies have shown an association between *FTO*, the fat-mass and obesity associated gene, and adulthood body-mass index (BMI). More thorough analyses utilising phenotypic data from several time points during the life course may deepen our understanding of the interplay between genetic and

environmental factors associated with BMI. We used structural equation modelling (SEM) technique to explore the network of variables associated with BMI from prenatal period until the age of 31 years.
The study sample consisted of individuals from the Northern Finland Birth Cohort 1966 (NFBC 1966), representing a population homogeneous for genetic, lifestyle and environmental factors. Of the original cohort, 4435 individuals had data on BMI at 31 years and had the variant rs9939609 in the FTO gene genotyped. In addition, data on maternal factors before and during pregnancy, outcomes related to birth, and aspects of life-style and health behaviour at 14 and 31 years were available.
Estimation of the effect of the FTO risk allele on the adulthood BMI, taking into account several relevant variables during the life course, showed evidence for association between FTO and BMI at the age of 31 years. An association with BMI at the age of 14 was also observed confirming the previous findings that FTO begins affecting well before adulthood. Weak evidence for association was observed between FTO and maternal BMI, and FTO and birth BMI. However, FTO explains only a minimal part of the trait variability, and environmental factors play a major role. We were not able to detect any interactions between FTO and other factors affecting BMI, possibly due to low power.

Keywords: Body-mass index, FTO, structural equation modelling.

Prognostic factors for short- and long-term graft survival in kidney transplantation in Estonia

Ü. Kirsimägi[1], **A. Lõhmus**[2], **P. Dmitriev**[2], **J. Kahu**[2], **A. Peetsalu**[1]
[1] University of Tartu, Estonia
[2] Tartu University Hospital, Estonia

Background. Forty years have passed since the first kidney transplantation in Estonia. The total number of kidney transplantations performed during 1968–2008 is 769. The average number of operations in the last 10 years was about 30–40 per million per year.
Aim of the study. The principal aim of the present analysis was to establish prognostic factors for short- and long-term graft survival.
Patients and methods. The analysis was based on 403 first cadaveric kidney transplantations in adult patients, carried out in Estonia in 1995–2007. The minimum follow-up period was at least 1 year, 60.3% of the patients were censored. Compared with the period 1995–1999, the results of graft survival were significantly better for the period 2000–2007 when the treatment scheme was supplemented with mono- and polyclonal antibodies for preventing rejection reaction. As 11.7% of the patients lost their graft during 3 postoperative months, the covariates affecting short-term graft survival need not be the same as the covariates important for the long-term period. Two different models, stratified for the two periods, were developed to estimate risk factors for graft failure for the short-term post-transplantation period (follow-up 1 year) and for the long-term period (follow-up 8 years) for the grafts that had survived the first three months (353 cases).
Results. The overall Kaplan-Meier estimate was 83.6 for one-year and 64.3 for five-year graft survival. Considering the whole study period, significantly better results were noted for the period 2000–2007 compared with the period 1995–1999, 1-year graft survival being 88.0% and 73.3%, respectively ($p = 0.0001$). Cox regression analysis revealed that the significant pre-transplantation characteristics affecting short-term graft survival are recipient age ≥ 50 years ($p = 0.003$) and body-mass index ≥ 30 ($p = 0.0103$), while the main predictor of graft survival is the post-transplantation characteristic delayed graft function ($p < 0.0001$). These prognostic factors for short-term graft survival do not significantly influence long-term graft survival for which the main prognostic factor is donor age ≥ 60 years ($p = 0.0003$).
Conclusions. Short-term graft survival is predominantly affected by delayed graft function, recipient age and body-mass index while long-term graft survival is affected by donor age. However it is obvious that longer follow-up of graft survival might yield additional risk factors influencing long-term graft survival.

Estimating Tree Survival on the Estonian Forest Research Plots Network

Allan Sims[1], **Maris Hordo**[1], **Diana Laarmann**[1]
[1] Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Tartu, Estonia

Tree survival, as affected by tree and stand variables, was studied using the Estonian database of permanent forest research plots. The tree survival was examined on the basis of remeasurements during the period 1995—2004, covering the most common forest types and all age groups. In this study, the influence of 35 tree and stand variables on tree survival probability was analyzed using the data of 31097 trees from 236 research plots. For estimating individual tree survival probability, a logistic model using the logit-transformation was applied. Tree relative height had the greatest effect on tree survival. However, different factors were included into the logistic model for different development stages: tree relative height, tree relative diameter, relative basal area of larger trees and relative sparsity of a stand for young stands; tree relative height, relative basal area of larger trees and stand density for middle-aged and maturing stands; and tree relative height and stand density for mature and overmature stands. The models can be used as preliminary sub-components for elaboration of a new individual tree based growth simulator.

Keywords: forest growth, logistic regression, generalized linear mixed model, survival probability, tree and stand variables

Bayesian modeling of recombination events in bacterial populations

Pekka Marttinen[1], **Adam Baldwin**[2], **William P. Hanage**[3], **Chris Dowson**[2], **Eshwar Mahenthiralingam**[4] **and Jukka Corander**[5]
[1]Department of Computational Science and Biomedical Engineering, Helsinki University of Technology, Finland
[2]Department of Biological Sciences, Warwick University, UK
[3]Department of Infectious Disease Epidemiology, Imperial College, UK
[4]Cardiff School of Biosciences, Cardiff University, UK
[5]Department of Mathematics, Abo Akademi University, Finland

We consider the discovery of recombinant segments jointly with their origins within multilocus DNA sequences from bacteria representing heterogeneous populations of fairly closely related species. The currently available methods for recombination detection capable of probabilistic characterization of uncertainty have a limited applicability in practice as the number of strains in a data set increases.

We introduce a Bayesian spatial structural model representing the continuum of origins over sites within the observed sequences, including a probabilistic characterization of uncertainty related to the origin of any particular site. To enable a statistically accurate and practically feasible approach to the analysis of large-scale data sets representing a single genus, we have developed a novel software tool (BRAT, Bayesian Recombination Tracker) implementing the model and the corresponding learning algorithm, which is capable of identifying the posterior optimal structure and to estimate the marginal posterior probabilities of putative origins over the sites.

A multitude of challenging simulation scenarios and an analysis of real data from seven housekeeping genes of 120 strains of genus *Burkholderia* are used to illustrate the possibilities offered by our approach. The software is freely available for download at URL `http://web.abo.fi/fak/mnf/mate/jc/software/brat.html`.

Keywords: bacterial recombination, bayesian modeling.

References:

Marttinen, P., Baldwin, A., Hanage, W.P., Dowson, C., Mahenthiralingam, E., and Corander, J. (2008). Bayesian modeling of recombination events in bacterial populations. *BMC Bioinformatics 9*, 421.

Bias in Odds Ratios by Logistic Regression Modelling and Sample Size

Szilárd Nemes[1] **Junmei Miao Jonasson**[1], **Anna Genell**[1] **and Gunnar Steineck**[1,2]
[1] Division for Clinical Cancer Epidemiology, Sahlgrenska Academy, University of Gothenburg, Sweden
[2] Division of Clinical Cancer Epidemiology, Department of Oncology and Pathology, Karolinska Institutet, Sweden

It is known that the estimated logistic regression coefficients are systematically biased given a small or moderate sample size.
The asymptotic bias of a maximum likelihood estimator, $bias(\beta)$, can be summarized as

$$bias(\beta) = \frac{b_1(\beta)}{n^1} + \frac{b_2(\beta)}{n^2} + ...$$

where $b_i(\beta)$ depends only on estimate beta coefficient, $\hat{\beta}$, but not of the sample size.
With help of a simulation study we demonstrate how the sample size determines the size of bias. We created a target population with 100000 individuals and from this draw repeated samples with $a\ priori$ determined sample sizes that varied from 100 to 1500 with increment 5. Then we fitted a least squares regression model to estimate $b_1(\beta)$, $\hat{\beta} = \beta_{pop} + b_1(\beta)n^{-1}$. As $n \to \infty$ the bias converges to zero $(b_1(\beta)n^{-1} \to 0)$, thus the intercept corresponds to unbiased estimate of the population parameter value. We concluded that studies employing logistic regression as analytical tool to study the association of exposure variables and the outcome overestimate the effect in studies with small to moderate samples size. This bias might in a single study not have any relevance for the interpretation of the results since it is much lower than the standard error of the estimate. But if a number of small studies with systematically overestimated effect sizes are pooled together without consideration of this effect we may misinterpret evidence in the literature for an effect when in the reality such does not exist.

Keywords: odds ratio, maximum likelihood, logistic regression, bias

References:

Pawitan Y (2001). *In all Likelihood. Statistical Modelling and Inference Using Likelihood.* Oxford Science Publications.

Income-related inequality in the utilisation of health care in Estonia

Janek Saluse[1] **and Andres Võrk**[1]
[1] University of Tartu, Estonia

Aim. The study assesses the income-related equity and inequality of health care utilisation in Estonia, a typical Eastern European country with high income inequality and high share of out-of-pocket payments in total health expenditures (one fifth in Estonia). This is one of the few attempts in assessing income-related equity and equality in the utilisation of health care in Eastern Europe using internationally comparative approach.
Methods. We use the methods outlined in Van Doorslaer, Masseria et al (2004), a major study on OECD countries. Income-related inequality in health care utilisation is measured with the concentration index (CI) and horizontal inequity is measured with the horizontal inequity index (HI) derived from CI after standardizing the need for health care. The income-related inequality is further decomposed into socio-economic determinants using a regression-based approach.
Data. The study uses micro level data from the Estonian Household Budget Survey 2006, which included a separate section on health status and health care utilisation in that year. Self-reported number of contacts with health service providers during last six months is used. We differentiate between visits to family doctor, dentist and other medical specialist, phone consultation, use of emergency medical care, day treatment and hospitalization. Need for health care is approximated with age, gender, self-reported health status and self-reported disability status. In total data on 7,826 individuals in 3,628 households were used.

Results. The results show negative income-related inequality for visits to family doctor, utilisation of emergency medical care and hospitalisation. After standardisation for health care need, these effects disappear and the utilisation of dental care, phone consultations and day treatment service turn out to be positively related to income.
Conclusions. In general, the results on Estonia are comparable to the results of OECD countries. The problematic areas of inequity are in the utilisation of dental care, phone consultations and day treatment service where the wealthier population has an advantage. The Estonian results imply that health care financing affects significantly the inequity in health care utilisation. Dental care for adults is not covered by the health insurance in Estonia and day treatment with high out-of-pocket payments is often used because of long queues in health insurance system.

Keywords: income-related inequality, health care utilisation, Estonia

---

SURVIVAL ANALYSIS OF RECTAL CANCER PATIENTS IN FINLAND USING PARAMETRIC MIXTURE MODELS

Teija Seppänen

University of Oulu, Finland

A common aim of population-based cancer survival analysis is the estimation of net survival, a measure of patient survival corrected for the effect of other causes of death. Net survival can be estimated using relative survival, that is the ratio of the observed survival in the patient group divided by the expected survival of a comparison group, assumed to be practically free of cancer of interest.
Patient survival for cancers of colon and rectum (CRC) have improved over the past few decades. CRC is the third most common cancer in Finland. Annually over 900 new rectal cancer cases are diagnosed and the disease causes about 400 deaths per year in Finland.
In this study the relative survival of rectal cancer patients in Finland is analysed. The actuarial life table method is used to describe the observed survival by gender, age, year of diagnosis and stage. The expected survival based on national life tables is estimated using the Hakulinen method. In addition we apply a parametric mixture model which provides a way of analysing simultaneously the proportion of patients cured together with the expected survival time of fatal cases. The survival time of the uncured is assumed to have the three-parameter generalized gamma (GG) distribution. The GG distribution contains the most commonly used distributions, including the exponential, Weibull and log-normal, as special cases. The BFGS algorithm implemented in R is used in the maximum likelihood fitting of these models.

Keywords: Survival analysis, Relative survival, Rectal cancer, Parametric mixture model.

References:

Cox, C., H. Chu, M.F. Schneider, A. Muñoz (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine 26*, 4352–4374.

De Angelis, R., R. Capocaccia, T. Hakulinen, B. Soderman, A. Verdecchia (1999). Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine 18*, 441–454.

Dickman, P.W., A. Sloggett, M. Hills, T. Hakulinen (2004). Regression models for relative survival. *Statistics in Medicine 23*, 51–64.

Estéve, J., E. Benhamou, L. Raymond (1994). *Statistical Methods in Cancer Research Volume IV: Descriptive Epidemiology*. IARC Scientific Publications No. 128. Lyon: International Agency for Research on Cancer.

ANALYSIS OF OVERDIAGNOSIS OCCURRING IN PROSTATE CANCER INCIDENCE IN FINLAND

Tiina Seppänen

University of Oulu, Finland

Prostate cancer is nowadays the most common cancer diagnosed among men in Finland. Its incidence has continuously increased during the past decades. The upward trend became steeper around 1990, which coincided with increasing use of opportunistic PSA testing since then.

By overdiagnosis of cancer we mean diagnosis of small slow-growing tumours that would never develop into a clinical phase and cause symptoms or premature death. Overdiagnosis can be a serious problem associated with PSA screening because it leads to overtreatment. Even though some evidence exists that PSA screening decreases mortality from prostate cancer, it is not yet recommended as routine screening.

In this study we attempt to estimate indirectly the extent of overdiagnosis of prostate cancer in Finland since 1990. The incidence of prostate cancer for men 50 years and older up to 2004 was modelled and predicted by an age-period interaction model using Poisson regression. Great differences implying hundreds of excess cases since 1990 were found between the observed and expected prostate cancer rates, when the latter were based on a prediction model continuing the long-term trend observed up to 1990. Some fraction of this excess is probably not representing overdiagnosis but clinically detectable cases brought forward in time by virtue of PSA testing.

Since 2004 the annual rates of prostate cancer seem to have turned down, thus following a similar trend as in the United States that started 10 years earlier. We shall construct predictions on the future incidence in Finland partly based on the US experience.

Keywords: Prostate cancer incidence, Overdiagnosis, PSA screening, Age-period interaction model.

References:

Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in Medicine 26*, 3018–3045.

Määttänen, L. (2007). *Performance of the Finnish Prostate Cancer Screening Trial Based on Process Indicators.* Doctorś thesis. University of Tampere: Acta Universitatis Tamperensis No. 1247.

Schröder, F.H., J. Hugosson et al. (2009). Screening and Prostate-Cancer Mortality in a Randomized European Study. *NEJM 360*, 1320–8.

Spiegelhalter, D.J. (1998). Bayesian graphical modelling: a case-study in monitoring health outcomes. *J. R. Statist. Soc. C, 47*, Part. 1, 115–133.

FORMIS - FOREST MODELING INFORMATION SYSTEM

Allan Sims

Department of Forest Management, Estonian University of Life Sciences, Tartu

Forest Modeling Information System (ForMIS) aims to systematize and structure existing information what we have for forest modeling and growth predicting in Baltic Sea region. The web based information system with combined use of empirical data and existing knowledge about forest development enhances the simplification of forestry metadata and forest models manipulation and verification/validation procedures.

In principle, information system contains two parts – empirical data management and models estimation. Data management includes data verification in data recording into the database, outliers detection after data recording and standard calculations like stand level data calculation from tree data. Knowing common sentence “all models are wrong, but some of them are useful” leads to model validation before using foreign models and also comparing different own country models.

Information system is available on the website `http://formis.emu.ee/`, where are presented empirical dendrometric equations, empirical datasets, growth and yield tables and list of theoretical forest equations. Information system is useful for forest modellers, who use empirical datasets, growth and yield tables and theoretical forest equations; and useful for practical foresters, who use empirical dendrometric models for managing and predicting forest resources.

Keywords: growth modeling, empirical datasets

Ordination of floristic data from young deciduous forest plantations using Non-metric Multidimensional Scaling (NMDS)

Tea Soo[1], **Arvo Tullus**[1], **Hardi Tullus**[1], **Elle Roosaluste**[2]
[1] Institute of Forestry and Rural Engineering, Estonian University of Life Sciences, Tartu, Estonia
[2] Institute of Ecology and Earth Sciences, University of Tartu, Estonia

The understorey vascular plant cover and its relations with the environmental variables were studied in 7- to 9-yr-old commercial hybrid aspen (*Populus tremula* L. × P. *tremuloides* Michx.) and silver birch (*Betula pendula* Roth) plantations on abandoned agricultural sites with a different land use (grassland or crop field) and site preparation (whole-area ploughing or strip tillage) history. A total of 248 vegetation plots (2 × 2m) were established within 62 experimental areas; vegetation descriptions were compiled, concentrations of total N, extractable P and K, and pH of the soil humus layer were determined, and growth traits of the trees were recorded.

In order to investigate how environmental variables corresponded to the understorey vegetation, Non-metric Multidimensional Scaling (NMDS) followed by the vector fitting was performed with the community ecology package Vegan 1.15-1 for R (Oksanen *et al.* 2008). The same data were analysed also with Detrended Correspondence Analyses (Hill, 1979), however the results from NMDS provided more reliable results for interpretations.

The NMDS ordination of the studied vegetation plots was significantly affected by previous land use, intensity of site preparation and chemical soil properties, the impact from overstorey was yet weak in young plantations. The results are presented in detail in separate publications (Soo *et al.*, 2009a, 2009b).

The study indicated that NMDS is an effective tool for ordination analyses of the understorey vegetation in young forest plantations on abandoned agricultural sites where the distinction of developing plant communities is still at an early stage.

Keywords: biodiversity, non-metric multidimensional scaling, plantation forestry, vegetation analyses

References:

Hill, M.O. (1979). DECORANA – a FORTRAN program for detrended correspondence analysis and reciprocal averaging. Cornell University, Ithaca, New York.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G.L., Solymos, P., Stevens, H.H., Wagner, H. (2008). The vegan Package. URL `http://cran.r-project.org/`, `http://vegan.r-forge.r-project.org/`.

Soo, T., Tullus, A., Tullus, H., Roosaluste, E. (2009a). Floristic diversity responses in young hybrid aspen plantations to land-use history and site preparation treatments. *Forest Ecology and Management*, *257*, 858–867.

Soo, T., Tullus, A., Tullus, H., Roosaluste, E., Vares, A. (2009b). Change from agriculture to forestry: floristic diversity in young fast-growing deciduous plantations on former agricultural land in Estonia. *Annales Botanici Fennici 46* (in press).

Spatial variations in alcohol related deceases in Finland: Case study with sparse Gaussian process models

Jarno Vanhatalo[1] **and Aki Vehtari**[1]
[1] Helsinki University of Technology, Finland

In this work we describe a detailed statistical analysis concerning the spatial variations of alcohol related diseases in Finland during 2001-2005. There are two objectives in the work. First to demonstrate a statistical analysis with a multiscale spatial prior that is constructed using sparse Gaussian processes for spatially very accurate data set, and secondly, to describe results that are important for health care authorities.

The mortality data used in the work is aggregated into a lattice of 5km×5km cells and contains approximately 7900 death cases during the five year period. The high resolution of the spatial data makes possible to analyze simultaneously both country wide and local phenomena. We will show that using multiscale model improves the results significantly compared to simple model with only one spatial component. The number of death cases during the five year period is so small (in average less than one in each cell) that a Poisson observation model is not sufficient. Thus we will use negative binomial distribution to allow over-dispersion. The total number of data points, approximately 10 500, is prohibitive for a naive Gaussian process implementation, and thus we will use sparse Gaussian processes (Vanhatalo and Vehtari, 2009).

Alcohol related diseases are considered one of the most severe problems in public health in Finland. These problems are not distributed evenly across the country, but some areas suffer the problem more than others. For this reason, a disease mapping study concerning the spatial variations of alcohol related diseases is needed to find causes for the problem. Thus, the results presented are of central importance in public health management. There are two findings in the data set: The alcohol related diseases spread unevenly between western and eastern Finland. The risk is lowest in west coast and highest in south-east of Finland. The other result is that the problems related to alcohol seem to concentrate in centers of population, such as cities and towns.

Keywords: disease mapping, sparse Gaussian process, alcohol related deceases, Finland, multiscale spatial model

References:

Vanhatalo J. and Vehtari A. (2009). Approximate Inference for Disease Mapping With Sparse Gaussian Processes, *Submitted*

Author index