

21st Nordic Conference on Mathematical Statistics

June 11 - 15, 2006

Comwell Rebild Bakker
Rebildvej 36, Rebild
DK-9520 Skørping
Denmark

Contents

The Scandinavian Journal of Statistics Prize	2
Conference Committees	3
Social Programme	3
Venue	4
Internet	4
Programme	4
Abstracts	5
Sunday, June 11	5
Monday, June 12	5
Tuesday, June 13	15
Wednesday, June 14	18
Thursday, June 15	28
List of Participants	33

21st Nordic Conference on Mathematical Statistics

June 11 - 15, 2006

The organizing committee welcomes you to the 21st Nordic Conference on Mathematical Statistics. The Nordic Conference on Mathematical Statistics (NordStat) is a biennial meeting for statisticians and probabilists from the countries in Northern Europe. NordStat also welcomes participants from countries outside Scandinavia, in particular from our close neighbours in the Baltic states. The official language at all sessions of this international conference is English and focus is on recent research in the Nordic and Baltic countries.

The meeting has several invited speakers - please see the programme - as well as a number of contributed sessions:

- Statistical Modelling of Time Series Data in Finance and Economics
- Infectious Epidemiology
- E-learning and Statistics
- Modelling using SDEs
- Graphical Models
- Statistics and scientific responsibility
- Spatial and Spatio-temporal Modelling
- Applied Probability (slet denne hvis ikke Anwar Joarder melder sig til)
- Biostatistics
- Statistics in Information Engineering
- Chemometrics
- Simulation-based Inference and Computational Statistics
- Statistical Analysis of Complex Event History Data
- Bioinformatics and Statistical Genetics

Once again the organizing committee welcomes you to the 21st Nordic Conference on Mathematical Statistics and wish you a pleasant and stimulating conference.

The Scandinavian Journal of Statistics Prize

The SJS Prize is to be awarded for the best paper submitted to Scandinavian Journal of Statistics and presented by a young statistician in the conference. This paper is to be published in a forthcoming issue of Scandinavian Journal of Statistics.

The Scandinavian Journal of Statistics Prize of 40.000 DKK will be given to the young researcher who presents the best paper at the NordStat meeting to be held in Rebild Bakker, Denmark, 2006, provided that the paper is accepted (before or after the meeting) as a regular submission to the Scandinavian Journal of Statistics (SJS). The editor of SJS must be informed that the paper is a candidate for the Scandinavian Journal of Statistics Prize.

To qualify as a candidate for the SJS Prize, the author must have received the Ph.D.-degree at most five years before the NordStat meeting or be younger than 35 years of age at the time of the meeting. The young author is preferably the sole-author of the paper, but if multiple young authors are involved, the paper will also be considered. There are no restrictions on nationality.

Young researchers of any nationality are strongly encouraged to present their work at the NordStat 2006 meeting and to submit their papers to the SJS and for the SJS Prize competition.

At the NordStat meeting in Stockholm 2002 the SJS Prize was given to Helle Sørensen for her outstanding paper on Simulated likelihood approximations for stochastic volatility models and her excellent presentation at the NordStat meeting in Grimstad, Norway 2000. The paper was published in the Scandinavian Journal of Statistics volume 30.

Conference Committees

The conference programme committee consists of

- Per Bruun Brockhoff (chairman)
Technical University of Denmark, Denmark
- Søren Asmussen
University of Aarhus, Denmark
- Bjarne Ersbøl
Technical University of Denmark, Denmark
- Jens Ledet Jensen
University of Aarhus, Denmark
- Niels Keiding
University of Copenhagen, Denmark
- Jesper Møller
Aalborg University, Denmark
- Tue Tjur
Copenhagen Business School, Denmark

The conference organizing committee consists of

- Kim Emil Andersen (chairman)
Vestas Asia Pacific, Denmark
- Per Bruun Brockhoff
Technical University of Denmark, Denmark
- Judith L. Jacobsen
Statcon ApS, Denmark
- Erik Parner
University of Aarhus, Denmark
- Jørgen Holm Petersen
University of Copenhagen, Denmark
- Helle Sørensen
The Royal Veterinary and Agricultural University, Denmark

Social programme

The opening ceremony will take place Sunday June 11, 2006 at 15:00 in the Large Auditorium of the hotel; see the enclosed map for directions. All evenings and the afternoon of Tuesday June 13 are to your free disposal. There is an optional excursion to the beautiful surroundings of Silkeborg. A few seats are still available.

The conference dinner will be held Wednesday evening.

Venue

The NordStat 2006 conference takes place on the top of one of the Danish National Parks and at the foot of Rold Forest (the largest forest in Denmark). The hotel offers a wealth of facilities both in- and outdoor. Go for a walk in Rold Forest and enjoy the magnificent nature. Or sweat in the fitness room and go for a swim in the pool.

Ask at the reception for maps, etc.

Internet

The hotel has Wireless High Speed Internet Access. Ask for access card in the reception. Moreover, computers with internet access will be provided by the conference.

Programme

See the enclosed programme for details. Abstracts are provided below.

Abstracts

Sunday, June 11

Statistical issues in determining cis-regulatory modules of transcription factors

Terry Speed (terry@stat.berkeley.edu)
University of California, Berkeley, USA

The lectures will outline some of the statistical methods used and challenges ahead as we try to help identify cis-regulatory modules. Loosely speaking, these involve the analysis of a range of types of data, both separately and jointly, with the aim of identifying individual and later sets of DNA sequence motifs which are involved in the regulation of the expression of particular genes in given tissues under given conditions. The types of data involved include genome sequence data from the organism of interest, and perhaps from related organisms, gene expression data, typically from microarrays, data on the likely location of transcription factor binding sites, typically from chromatin immunoprecipitation (ChIP) assays, with subsequent sequencing or hybridizing to microarrays, and perhaps gene disruption assays. The statistical analyses currently range from motif search algorithms, linear and nonlinear regression with thousands of variables, change-point like algorithms that seek rare events along DNA sequences, methods for finding association between two or more point processes, and more.

Monitoring performance in the UK health-care system: the role of statistical methods

David Spiegelhalter (davids@mrc-bsu.cam.ac.uk)
MRC Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom

A combination of major health scandals, and a government obsession with target-setting as a means of improving public services, has led to a strong move towards a system for routine monitoring of health-care performance in the UK. A controversial 'star-ratings' approach for assessing all National Health Service Trusts is due for a major overhaul by the Healthcare Commission, which is the body responsible for inspection and quality improvement. Statistical analysis of performance indicators will form a substantial component of any future monitoring and assessment procedure, and issues such as 'league tables', methods for comparing performance against targets, and combination of multiple indicators into summary measures are all being discussed. Adaptations of classical industrial quality-control procedures, such as over-dispersed funnel plots and risk-adjusted CUSUMS, are being explored as possible tools to help in comparing performance and guiding inspections in this high-profile and somewhat politicised area. The application of such methods to routinely collected data will be illustrated.

Monday, June 12

Asymptotic inference for the AR-ARCH model

Rasmus Theis Lange (lange@math.ku.dk)
Department of Applied Mathematics and Statistics, University of Copenhagen, Denmark

We consider asymptotic inference in the Autoregressive (AR) model with ARCH innovations. We discuss consistency and asymptotic normality of the quasi-maximum likelihood estimator and a modified quasi-maximum likelihood estimator (based on censoring away extreme observations) under much less strict moment restrictions than the ones currently employed in the literature. Furthermore mild conditions for geometric ergodicity of the AR-ARCH process are derived. Finally the asymptotic properties of the estimators are examined by a Monte Carlo study and advice on practical estimation is provided.

Representation theory for a class of vector autoregressive models for fractionally integrated processes

Søren Johansen (sjo@math.ku.dk)

Department of Applied Mathematics and Statistics, University of Copenhagen, Denmark

Based on an idea of Granger (1986), we analyse a new vector autoregressive model defined from the fractional lag operator $1 - (1 - L)^d$. We first derive conditions in terms of the coefficients for the model to generate processes which are fractional of order zero. We then show that if there is a unit root, the model generates a fractional process X_t of order d , for which there are vectors β so that $\beta'X_t$ is fractional of order $d - b$.

We find a representation of the solution which demonstrates the fractional properties. Finally we suggest a model that allows for a polynomial fractional vector, that is, the process X_t is fractional of order d , $\beta'X_t$ is fractional of order $d - b$ and a linear combination of $\beta'X_t$ and $(1 - L)^b X_t$ is fractional of order $d - 2b$. The representations and conditions are analogous to the well known conditions for $I(0)$, $I(1)$ and $I(2)$ variables.

The likelihood ratio test for cointegration ranks in the $I(2)$ model

Heino Bohn Nielsen (Heino.Bohn.Nielsen@econ.ku.dk)

Department of Economics, University of Copenhagen, Denmark

The likelihood ratio test for cointegration ranks in the $I(2)$ model This paper presents the likelihood ratio (LR) test for the number of cointegrating and multi-cointegrating relations in the $I(2)$ vector autoregressive model. It is shown that the asymptotic distribution of the LR test for the cointegration ranks is identical to the asymptotic distribution of the much applied test statistic based on the Two-Step estimation procedure in Johansen (1995), Paruolo (1996), and Rahbek, Kongsted, and Jørgensen (1999).

By construction the LR test statistic is smaller than the non-LR test statistic from the Two-Step procedure, and application of the LR test may change rank selection in empirical work. Based on a study of existing empirical applications and related Monte Carlo simulations we conclude that the LR test has much better size properties when compared to the Two-Step based test. Overall, we propose to use of the LR test for rank determination in empirical applications of the $I(2)$ model.

Random graphs, infectious diseases and vaccination

Tom Britton (tom.britton@math.su.se)

Department of Mathematics, Stockholm University, Sweden

Random graphs of various types can be used to describe the social structures in a community. Given such a graph a simple epidemic model can be defined in which individuals may infect a subgroup of its acquaintances. In order to avoid a major epidemic outbreak the community might be partly vaccinated. In the present talk we study these types of problems with focus on graphs where the degree distribution is heavy-tailed but other than that completely random. For such graphs we derive results determining under what criteria a major outbreak may occur, and for what vaccination policies such outbreaks are avoided

Endemic persistence or disease extinction: the effect of population separation into subcommunities

Mathias Lindholm (lindholm@math.su.se)

Department of Mathematics, Stockholm University, Sweden

Suppose that an infectious disease which is endemic in a population divided into several large subcommunities which

interact. Our aim is to understand how the time to extinction is affected by the interaction between communities.

We present two approximations for the expected time to extinction in a symmetric population structure consisting of a small number of large communities. These approximations are derived for an SIR model with demography with focus on diseases with short infectious period in relation to life length, such as childhood diseases and influenza. Both approximations are based on Markov jump processes.

Simulations indicate that the time to extinction is increasing in terms of interaction between communities. This behaviour can also be seen in our approximations in regions of the parameter space corresponding to diseases like influenza.

Genetic and environmental influences on birth weight, birth length, head circumference and gestational age using population based parent-offspring data

Astrid Lunde, Kari Klungsøyr Melve, Hkon K. Gjessing, Rolv Skjærven
and Lorentz M. Irgens (astrid.lunde@mfr.uib.no)

Section of Epidemiology and Medical Statistics, Department of Public Health and Primary Health Care, University of Bergen, Norway

Familial correlations in birth weight and gestational age have been explained by fetal and maternal genetic factors, mainly in studies of offspring of twins. The aim of the present intergenerational study was to estimate and compare fetal and maternal genetic effects and shared sibling environmental effects on birth weight and gestational age, but also on crown-heel length and head circumference. We used path analysis and maximum likelihood models to estimate these effects, and at the same time adjust for covariates. Parent-offspring data were obtained from the Medical Birth Registry of Norway (MBRN) from 1967 to 2004. For the analysis of birth weight and crown-heel length, 101,748 families were included, for gestational age 91,617 families, and for head circumference, 77,044 families were included. Fetal genetic factors explained 31% of the normal variation in birth weight and birth length, 27% in head circumference and 11% in gestational age. Maternal genetic factors explained 22% of the variation in birth weight, 19% in birth length and head circumference, and 14% in gestational age. Relative to the proportion of explained variation, fetal genes were most important for birth length and head circumference.

Master of Applied Statistics – A Web-Based Statistics Degree

Bent Jørgensen (bentj@stat.sdu.dk)

Department of Statistics, University of Southern Denmark, Denmark

Master of Applied Statistics (MAS) is a flexible web-based adult education program in English, designed to meet the demands for training in modern applied statistics methods. MAS is a 2 1/2 year part-time program, which started in 2003, so the first students have just finished. MAS is offered by the University of Southern Denmark, in collaboration with the Royal Veterinary and Agricultural University of Denmark, the Technical University of Denmark and the Danish Institute of Agricultural Sciences.

Many company employees responsible for data handling or statistical analyses find themselves in the situation, that their educational background in statistics is limited, or dates back a long time. Some may have acquired good skills at employing certain statistical methods, but there are often doubts - do I use the best and most correct method for this data or that analysis, or might there be problems with the methods I normally use? Are there new methods available, that I do not know about? What could I do better if I was in command of the newest software in the area? In short, one may lack an overview of suitable statistical methods, and knowledge about when and how to apply these methods in practice.

Employees, who find themselves in such a situation have probably considered the possibility of taking further education in statistics. Some may have participated in short courses, but may still feel it difficult to achieve a good enough overview of the field to be able to apply statistical methods with confidence.

Those who contemplate participating in longer and more comprehensive educational programs in statistics or similar areas, often face insurmountable practical problems, because traditional educational programs are more often than not incompatible with the demands of juggling job, family and leisure, in particular if the education is not situated in one's own city.

The MAS program aims at giving the participants an overview of modern statistical methods, and enable them to perform statistical analyses using modern statistical software packages. Except for one kick-off meeting per course, the instruction is entirely web-based, and uses the Blackboard learning management system.

In this way, MAS is able to meet the demands of most companies and employees for further education in statistics today. The use of e-learning is crucial for our students, who tend to be full-time employees in private or public companies. In the talk, I will give an overview of MAS, and mention some of the problems in e-learning that we face. Further information (in Danish) about MAS is available at the site statmaster.sdu.dk.

E-learning in Practice

Pia Veldt Larsen (p.v.larsen@stat.sdu.dk)

Department of Statistics, University of Southern Denmark, Denmark

New possibilities and new challenges arise as e-learning is making its entry into teaching. In this talk, considerations and experiences about practical, technical and pedagogical issues regarding e-learning will be presented.

Flexibility is an obvious advantage of e-learning. The electronic medium allows students to participate in the course when and where it is most convenient for them. Distribution of course material, links, notes, announcements, etc. is quick and easy, and all online-discussions are stored and can, at any time, be re-read and resumed. Linking discussions from different modules, for example by making references to previous (or later) discussions, can help the students to gain a better overview and sense of coherence of the course. Also, an active online discussion forum, in which students discuss the course material and help each other, can be very stimulating for the learning process.

One of the major challenges of e-learning is the lack of 'non-verbal' communication – eye-contact, body language, intonation, phrasing, etc., not to mention the use of drawings and sketches to explain and exemplify difficult steps in the theory. This can cause misunderstandings and frustrations for students as well as teachers: How to formulate a question in writing on something one does not understand? How to answer a 'nonsense' question that uses ambiguous notation and terminology, or simply is unintelligible? And how to write an answer in a unambiguous way and at the appropriate level of the students?

Other challenges include replacing lectures as a mean to structuring the course and providing overview and coherence; activating the students during the course; keeping the focus on statistics and not get sidetracked by unimportant or irrelevant details – or by (computer-)technical difficulties.

The talk will be illustrated by examples from a statistics course in the webbased Master of Applied Statistics programme.

Learning Objects - a New Way of Teaching Statistics

Helle Rootzén (hero@imm.dtu.dk)

Informatics and Mathematical Modelling, Technical University of Denmark

Students nowadays differ much more than before - some are very good and some have substantial difficulties even with very basic concepts - "Learning Objects" may be used to ensure that each student gets course material which is at the right level. Further students have very different learning styles - some learn best by first getting exposed to theory and afterwards seeing examples, while others prefer the opposite order of presentation. Similarly some prefer visual and graphical teaching, some like to see theory written down in formulas, while others get the most out of listening to oral presentations - "Learning Objects" make it possible for each student to use what suits her best. "Learning Objects" represent a relatively new method of subdividing courses into smaller modules. According to [1], a "Learning Object" is defined as follows: "Any digital resource that can be reused to support learning. The term "Learning Objects" generally applies to educational materials designed and created in small chunks for the purpose of maximizing the number of learning situations in which the resource can be utilized".

Usual learning - by eg using a book - is linear. You start at page 1 and continue on - hopefully understanding a little bit all the time. To learn in this way is one example of a learning style among many - experience and theory suggest that different people learn in different ways. Another way of learning - which is not supported by reading a book - is by getting a lot of information and then suddenly understanding the whole. For a description of different learning styles see [3]. Books and lectures can be good instruments for learning but should not stand alone. To combine them with a more "anarchistic" kind of material can improve learning by making it more individual and more fun. With "Learning Objects" you can build your own course to suit your learning style and to provide you

with optimal learning.

Faced with a new generation of students who are used to exploiting the possibilities of the computer, we need a new type of education that will reflect a rethinking of content, form and duration. In the future, education will be in the form of "voucher systems". You get a set of vouchers and use them to attend the specific chunk of a study programme you need whenever and wherever it suits you. If the providers are to meet these requirements, the task of developing new courses and tailoring these to new students must be manageable. We therefore propose a new type of courses. These are structured around "Learning Objects", short complete education sessions, which may be combined in various ways according to the student's interests and levels. We combine the "Learning Objects" with "blended learning" and the ideas are tried out in research-based continuing education in applied statistics. Working with "Learning Objects" gives a wide range of flexibility for both the course providers and the users. E.g., the structure makes it easier to suit different learning styles and the reusability makes it easier to make new courses tailored for new customers.

An important part of our efforts is to create sufficient computer support for cost-effective course development. This is done in three ways: Development of a dedicated repository system using metadata, [2], for easy storage and retrieval of "Learning Objects", creation of a course-making tool which combines learning objects into courses, and construction of a tool for making "Learning Objects".

References:

1. <http://wiley.ed.usu.edu/docs/encyc.pdf>
2. <http://ltsc.ieee.org/wgl2/index.html>
3. R.M. Felder and R. Brent, 'Understanding Student Differences'. J. Engr. Education, **94**(1), 57-72 (2005).

Hidden Markov and state space models – from likelihood theory to computational statistics

Tobias Rydén (email)

Centre for Mathematical Sciences, Lund University, Lund, Sweden

During the the last 10-15 years, hidden Markov and more general state space models have undergone a rapid development in terms of theoretical and computational statistics. On the inferential side most of the focus has been on likelihood theory, which in many respects today is rather complete – at least for models with compact state space.

However, except for models with finite state space and linear Gaussian models, the likelihood cannot be computed, let alone maximised, exactly. Many procedures to compute approximations have been proposed, with so-called sequential Monte Carlo methods, or particle filters, having received much attention recently.

The topic of this talk is the merge of these two areas: finding numerical methods that produce approximate ML estimators that at least asymptotically have favourable properties. We will talk about estimators derived from numerical approximations of the likelihood function, Monte Carlo EM algorithms, simulated annealing-type algorithms, and also about their performance in terms of consistency and efficiency.

Problem solving is often a matter of cooking up an appropriate Markov chain

Olle Häggström (olleh@math.chalmers.se)

Chalmers University of Technology, Gothenburg, Sweden

By means of a series of examples, taken from classic contributions to probability theory as well as from my own practice, I will try to convince the audience of the claim made in the title of the talk. Along the way, I will have reason to discuss topics such as coupling, correlation inequalities, and percolation.

Hermite series with given marginal distribution and autocorrelation function

Bo Markussen (bomar@kvl.dk)

Department of Natural Sciences, Royal Veterinary and Agricultural University, Denmark

A diffusion-type process $X(t)$ is constructed as an Hermite series in sums of Ornstein-Uhlenbeck processes, i.e. via the so-called chaos expansion. We are concerned with parameter estimation given the marginal distribution and autocorrelation function of an observed signal: An explicit method of choosing the coefficients in the chaos expansion to fit the empirical marginal distribution is devised. An explicit formula for the autocorrelation function for $X(t)$ is presented, allowing the parameters in the Ornstein-Uhlenbeck processes to be fitted by least squares estimation against the empirical autocorrelation function. Finally, we describe a method to perform model control. The proposed algorithms are exemplified by the investigation of a sound signal.

SDEs in Drug Development: Non-Linear Mixed-Effects Modeling Based on SDEs

Rune Viig Overgaard (rvo@imm.dtu.dk)

Informatics and Mathematical Modelling, Technical University of Denmark, Denmark

Stochastic Differential Equations (SDEs) could potentially aid many parts of PK/PD modelling by a more complete description of the variations, better simulation properties, as a diagnostic tool that can pinpoint model deficiencies, etc. These models offer a general intra-individual error structure, where the residuals are decomposed into system noise from the SDEs and uncorrelated measurement noise.

Over the past few years, increasing interest has emerged for SDEs in PK/PD modelling, and some selected projects shall be reviewed: 1) CTSM is a stand-alone software for estimation in SDEs with measurement noise, and has been used for individual PK/PD modelling, e.g. to deconvolute the rate of appearance for linear and non-linear disposition models¹. 2) The Extended Kalman Filter can be merged with the FOCE algorithm for estimation of non-linear mixed-effects models based on SDEs. MATLAB was used to confirm that inter-individual variability, measurement- and system noise can be separated, such that these models can be treated meaningfully². 3) The Extended Kalman Filter was implemented in NONMEM to facilitate SDEs in more general PK/PD models and to increase estimation speed and performance.

The implementation of SDEs in NONMEM has enabled more recent applications of SDEs in population PK/PD modelling, and we shall summarize three unpublished models for SDEs in PK/PD: 1) A model for thermoregulation, where SDEs offer a more complete description of the variations and better simulation properties, 2) A model for insulin secretion, where SDEs changes the individual estimates, and 3) A model for haemoglobin, where SDEs decrease the sensitivity to model misspecification and increase the predictive performance.

References:

1. environmental factors and polygenes describe a large number of potential background influences
2. Kristensen NR, Madsen H, Ingwersen SH. A Deconvolution Method for Linear and Nonlinear Systems based on Stochastic Differential Equations. PAGE poster presentation 2004.
3. Overgaard RV, Jonsson N, Tornoe CW, Madsen H. Non-linear mixed-effects models with stochastic differential equations: implementation of an estimation algorithm. *J.Pharmacokinet.Pharmacodyn.* 2005; **32**(1):85-107.
4. Tornoe CW, Overgaard RV, Agerso H, Nielsen HA, Madsen H, Jonsson EN. Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharm.Res.* 2005; **22**(8):1247-1258.

First passage time distribution for the Ornstein-Uhlenbeck process

Susanne Ditlevsen and Ove Ditlevsen (S.Ditlevsen@pubhealth.ku.dk)

Department of Biostatistics, University of Copenhagen, Denmark

One-dimensional diffusion processes have been widely used as models for the membrane potential of neurons if the variability in the neuronal activity should be described. An action potential or a spike (when the neuron fires) is produced when the membrane voltage reaches some threshold, and corresponds to the first passage time for the associated stochastic process describing the membrane potential. The voltage following a spike is reset to its initial value, and time intervals between action potentials are identified with experimentally observed interspike intervals. This model is important due to the generally accepted hypothesis that the information transferred within the nervous system is encoded by the timing of the action potentials. Among the stochastic neuronal models the Ornstein-Uhlenbeck process plays a central role because it is the simplest stochastic model that includes the spontaneous decay of the membrane potential to a resting value, and appears as a good compromise between tractability and realism of the model.

The observed interspike intervals can be used to infer about the parameters of the underlying process, and it is therefore of interest to describe the distribution of the first passage time. The exact distribution of the first passage time is only known in special situations, namely when the asymptotic mean of the Ornstein-Uhlenbeck process equals the threshold value, and in the limit when the membrane time constant goes to infinity. In the latter case the process approaches a Wiener process, and the first passage time distribution through a constant threshold follows the inverse Gaussian distribution.

In this talk we propose an explicit approximation to the first passage time distribution when the asymptotic mean of the Ornstein-Uhlenbeck process is equal to or larger than the threshold (suprathreshold regime). In the two above-mentioned limits the approximation is exact and, moreover, certain moments are exact. The approximation is evaluated by stepwise numerical integration of the Chapman-Kolmogorov integral equation, and correction factors are obtained hereby. The approximation allows a derivation of approximate maximum likelihood estimates of the parameters of the underlying Ornstein-Uhlenbeck process.

Are Option Values Stochastic? On Effects of State and Parameter Uncertainty

Erik Lindström (erikl@maths.lth.se)

Centre for Mathematical Sciences, Lund University, Sweden

In the talk we analyze option values in the Black & Scholes framework, when agents only have access to a finite sequence of observations. This is the case for all real world applications. It will be shown that option values predicted by the Black & Scholes formula will be stochastic and hence have to be treated as such. The reason for this is that agents cannot observe one of the parameters (the volatility) used in the valuation formula. Furthermore, different agents use different estimators and different sets of data to estimate the parameter. This adds to making the de facto market volatility stochastic, even in the Black & Scholes market. Additional theoretical support for stochastic option values is obtained when stochastic volatility (latent factor) models are introduced. The additional stochastic element is essentially the non-linear filtering problem, as the volatility is unobservable, while the standard option valuation framework explicitly assumes the latent volatility to be known. However, assuming that the agents are aware of this problem and that they are using the best possible projection of the stochastic values to values measurable with respect to the available information generates some interesting stylized facts on the volatility structure which are consistent with the observed option volatility structure.

Key words: Option pricing; parameter estimation; optimal filtering; bias correction; Bayesian analysis.

Free, cross-platform gRaphical software

Claus Dethlefsen (aas.claus.dethlefsen@nja.dk)

Aalborg Hospital, Aarhus University Hospital, Denmark

A graphical model is a class of statistical models that can be represented by a graph which can be used to identify conditional independence properties. Some common examples of graphical models are Bayesian networks (directed

graphical models), log-linear models (undirected models), block-recursive graphical models, and models defined using the BUGS language. Today, there exists a wide range of packages to support the analysis of data using graphical models. Here, we focus on Open Source software, making it possible to extend the functionality by integrating these packages into more general tools.

We will attempt to give an overview of the available Open Source software, with focus on the gR project. This project was launched in 2002 to make facilities in R for graphical modelling. Several R packages have been developed within the gR project both for display and analysis of graphical models. This facilitates extensions in the form of R packages which may rely on the whole R system.

Examples of R packages in the gR project include: gRbase, defining a common data structure; giRaph, defining mathematical graphs; dynamicGraph, giving an interactive graphical user interface for manipulating graphs; CoCo, a bundle of algorithms for efficient analysis of graphical models; mimR, giving an interface to the MIM program; deal, learning parameters of a Bayesian network; ggm, giving tools for Gaussian graphical models; BRugs, running BUGS within R; SIN, model selection in Gaussian graphical models.

Coloured Graphical Gaussian Models

Søren Højsgaard (sorenh@agrsci.dk)

Danish Institute of Agricultural Sciences, Research Center Foulum, Denmark

Undirected graphical Gaussian models restricts elements of the concentration (inverse covariance) matrix K to being zero whenever the associated variables are conditionally independent given the remaining.

We introduce coloured graphical models by partitioning the vertices of the graph in vertex colour classes and the edges in edge colour classes. Models of one type restrict elements in K to being identical if the corresponding vertices/edges are in the same colour class. Another type of models restrict partial correlations to being identical. A subset of these models are determined by invariance under special permutations.

The properties of the models and associated estimation algorithms are discussed and illustrated. This represents joint work with Steffen Lauritzen, University of Oxford.

Graphical chain models for the analysis of complex genetic diseases

Di Serio Clelia (mail missing)

affiliation

Modelling genetic diseases is in general a hard task. It is difficult to make inference on the expressions of a quantitative observed trait describing the disease (phenotype) starting from the genetic information (genotypes). Genetic diseases can be roughly divided into two main categories: Mendelian and multifactorial genetic disorders. Mendelian disorders are rare and mainly monogenic, meaning that the disease is due to a single gene mutation. The phenotype for these diseases can be clearly identified and the distinction between affected and unaffected population is clear-cut. Mutations are rare and recent so that a causal gene-disease transmission mechanism can be identified. Multifactorial genetic disorders (such as hypertension, multiple sclerosis, schizophrenia, diabetes and other common disorders) are far more common than Mendelian disorders, and there is no defined pattern of segregation in families. The difficulty of dealing with these disorders is due to the following factors: a) misleading definition of the phenotype, b) many genes and environmental factors are jointly involved, c) the low penetrance of the disease meaning that mutations in frequent alleles are rarely associated with the disease, so that a disease-cause mechanism cannot be assessed d) no clear cut-off between affected and unaffected population. These diseases are not transmitted but only promoted by a collection of factors some of which are hereditary.

In this contribution a new statistical perspective to approach the multifactorial genetic disorders is introduced consistently with the genetic representation: the search for the pathogenetic mechanism of a disease goes from the top of complexity, at whole organism level, down to the DNA level by dissecting the disease phenotype into several intermediate phenotypes at different levels of biological organisation. We propose graphical chain models as a natural and intuitive statistical tool to investigate these problems. In fact, the graphical representation of these models allows to read in statistical terms the top-down approach to a genotype-phenotype chain (from the complexity of the phenotype to the simple DNA sequence alteration) as it follows:

1. environmental factors and polygenes describe a large number of potential background influences
2. these influences are acting at different levels (sub-cellular/cellular/organic/whole body) determining a de-

composition of the main problems in a set of sub-problems

3. intermediate responses (intermediate phenotypes) influence the final responses conditioning on the potential background influences
4. mixed multivariate final responses can be identified (final phenotypes)

A recursive structure is used to decompose the original complex framework into a set of smaller contexts and a conditional independence structure can be designed. In other words the structural information relating one component to the others is entirely contained in the conditional distribution function whenever each component is described by a random variable.

Elaboration, explanation and specification in graphical models

Svend Kreiner (s.kreiner@biostat.ku.dk)

Department of Biostatistics, University of Copenhagen, Denmark

Estimation of conditional relationships is very important in quantitative social research where the analysis is sometimes described as a process of elaboration, explanation and specification. Graphical models are natural frames of inference for such analyses because the graphical structure of the model can be used to determine the variables that have to be taken into when the strength of the conditional relationship between two variables are estimated.

In cases where the graphical model is not completely specified by subject matter considerations we have to apply model search procedures before relationships are estimated. Model search strategies in such situation should be regarded as part of the estimation procedure rather than a separate model building procedure, implying among other things that conventional ways of assessing the properties of the estimates no longer apply. The degree to which this is a problem is illustrated by analysis of a 15-dimensional contingency table where the properties of estimates based on simple model search strategies are assessed by bootstrapping and compared to the asymptotic properties of estimates when the model generated by the model search procedure is taken at face value.

The Danish controversy on "Social heredity"

Gorm Gabrielsen (gg.mes@cbs.dk)

Copenhagen Business School, Copenhagen, Denmark

The term *Social Reproduction* covers broadly speaking the pattern of intergenerational transference of social deprivation. These patterns are known and studied in many cultural settings around the world. In 1967, however, Gustav Jonsson in his thesis "Delinquent boys, their parents and their grandparents" introduces the term *social heredity*. Jonsson uses the concept to "explain" the criminality of boys as an inheritance from their parents. According to his theory it is therefore possible from knowledge of the behaviour of the parents to predict the fortune of the children. In this interpretation the social heredity seems to be the key to explain or key to understanding of some peoples severe social problems.

The concept of *social heredity* has achieved an enormous popularity in general. In Denmark we can almost every day read the newspapers referring research results - being explained from *social heredity*. It looks like all kind of deviances from some kind of (socially) expected behaviour can be explained from *social heredity* - all kinds of *non-behaviour* or *non-fulfilling*. Non-health, non-employment, non-education, non-sufficient education, non-legal behaviour - or more general non-integration - even school children's non-healthy packed lunch can be explained from social heredity.

The concept of social heredity in some sense seems to imply that we think and analyse at the individual level. Why does this *specific* boy show a criminal behaviour? Why does this specific girl bring an unhealthy packed lunch? The concept encourages explaining on the individual level although the concept carries a duality: *social* (common or aggregated) *heredity* (individual). It is therefore not astonishing that a lot of research about social heredity is concerned with determining factors to predict unwanted behaviour - so called risk factors. Which individual/children are in risk situation? Even a name has in Denmark been developed for such children - "risk-children" - a term, which again implicitly refers to the level of individuals - to specific children.

To illustrate the points I will show some examples of statistical analyses where it is unclear on which level the analysis is performed. Furthermore, I will show how this uncertainty effects presentation of the result and how this may affect the level of political action.

Correction of mistakes in scientific publications?

Ole Olsen (ole.olsen@gpract.ku.dk)

Research unit of general practice, Copenhagen

Statisticians have a core role in the production of knowledge in modern society. We develop, apply, disseminate and teach sharp tools for analyses of data. We may even care about the validity of the results in the studies we are involved in.

But who cares about the validity of the totality of quantitative research? How many misleading studies are needed to make the overall production of knowledge misleading and meaningless? How difficult is it to verify and correct errors? And what is the sociology of errors and corrections?

I will present two case stories (1,2). During a project I discovered that three Cochrane reviews on three major perinatal interventions all had, at different points in time, included, excluded and possibly again included, one trial each, for suspicion of scientific misconduct. The rate of possible scientific misconduct seemed astonishingly high. For all three systematic reviews, the initial results of the suspected trials were very promising outliers and the statistical results of the meta-analyses were sensitive to inclusion or exclusion. The suspected trials had all been important for expensive policy decisions. Both the investigations of the unverifiable results were protracted and painful processes.

No formal denominator exist for my few case stories but one wonders if statistics is a too common form of lying.

References:

1. Olsen O, Gøtzsche P. Nødvendigheden af elektronisk opdaterede meta-analyser: Doppler-ultralyd i obstetrikken som eksempel. *Ugeskr Læger* 1996;159(1):27-8.
2. Olsen O. Correction of misleading data delayed for 16 years [letter]. *Lancet* 2001;357:1360-1.

Some reflections on statistics and scientific responsibility

Inge Henningsen (inge@math.ku.dk)

Department of Applied Mathematics and Statistics, University of Copenhagen

"Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cook book, believing the recipes will work without understanding why. A more CORDON BLEU attitude to the maths involved might lead to fewer statistical souffles failing to rise."

Sloppy Stats Shame Science, *The Economist* June 3rd 2004

A number of articles in scientific journals have documented the fact that scientific literature is marred by statistical mistakes. Moreover, statistics plays a major role in some of today's more controversial research, and a number of cases involving statistics have demonstrated how public opinion on some big and complicated issues is formed outside the established scientific world. This may be seen as part of a shift from a mainly university based research to a knowledge production situated in a broader context, where the distinctions between basic and applied research are blurred. This new knowledge production demands new rules for good scientific conduct and a system of quality control that is able to comprise the diversity of contemporary research. The fact that quantitative methods seem to be playing an ever increasing role in new and diverse settings is a challenge to statistics and to statisticians. What should be the role of the statistical community in this new scientific order? What are the challenges for the Nordic Statistical Societies? In his plenary lecture David Spiegelhalter talks on "Monitoring performance in the UK health-care system: the role of statistical methods". In this workshop two invited speakers from Denmark take on issues from the social and the medical sciences, but it is hoped that other participants will furnish their own cases as a starting point for a discussion on statistics and scientific responsibility.

Tuesday, June 13

Modern Statistics for Spatial Point Processes

Jesper Møller and Rasmus Waagepetersen (jm@math.aau.dk,rw@math.aau.dk)

Department of Mathematical Sciences, Aalborg University, Denmark

We summarize and discuss the current state of spatial point process theory and directions for future research, making an analogy with generalised linear models and random effect models, and illustrating the theory with various examples of applications. In particular, we consider Poisson, Gibbs, and Cox process models, diagnostic tools and model checking, Markov chain Monte Carlo algorithms, computational methods for likelihood-based inference, and quick non-likelihood approaches to inference.

Deterministic inference for log-Gaussian Cox processes

Hanne Wist Rognebakke and Håvard Rue (hanne.rognebakke@nr.no)

Norwegian Computing Center, Norway

Log-Gaussian Cox processes (LGCPs) are flexible models for aggregated spatial point patterns. In an LGCP the logarithm of the intensity surface is a Gaussian process. From an observed point pattern, inference for the model parameters in the underlying Gaussian field can be made in a Bayesian setting using MCMC simulation. However, the simulation will be time consuming when the dimension of the covariance matrix is large.

Deterministic inference can be used to obtain approximated posterior inference for Gaussian Markov random field (GMRF) models, where the computational cost is insignificant compared to MCMC calculation (see Rue and Martino, 2005). Inference can be made both for the model parameters and each separate node in the GMRF, conditioned on the data. These deterministic schemes can be used for Gaussian random field (GRF) models as well, but due to the full precision matrices, the dimension of the problem will be more restricted than for GMRF models with sparse precision matrices.

By using a GMRF as an alternative model for the logarithm of the intensity surface, we obtain a sparse precision matrix, which results in a significant speed up in the simulation algorithms. GMRFs give a good approximation to GRFs with commonly used covariance functions using only a small local neighborhood (see Rue and Tjelmeland, 2002).

We present results for GMRF models in LGCPs using deterministic inference as well as MCMC. Due to the difficulty of estimating both the precision and range parameter in the covariance function of the GMRF, we also discuss the use of intrinsic GMRF models.

References:

1. Rue, H. and Martino, S. (2005). Approximate inference for hierarchical Gaussian Markov random field models. Statistics Report No. 7, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.
2. Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. Scandinavian Journal of Statistics, **29**(1):31–49.

A spatio-temporal model for fMRI data - with a view to resting state networks

Thordis Linda Thorarinsdottir and Eva B. Vedel Jensen (disa@imf.au.dk)

The T.N. Thiele Centre, Department of Mathematical Sciences, University of Aarhus, Denmark

Functional magnetic resonance imaging (fMRI) is a technique for studying the active human brain. During the fMRI experiment, a sequence of MR images is obtained, where the brain is represented as a set of voxels. The data obtained are a realization of a complex spatio-temporal process with many sources of variation, both biological and

technical. Most current model-based methods of analysis are based on a two-step procedure. The initial step is a voxel-wise analysis of the temporal changes in the data while the spatial part of the modelling is done separately as a second step in the analysis. We present a spatio-temporal point process model approach for fMRI data where the temporal and spatial activation are modelled simultaneously. This modelling framework allows for more flexibility in the experimental design than most standard methods. It is also possible to analyze other characteristics of the data than just the locations of active brain regions, such as the interaction between the active regions. We discuss various methods for statistical inference in the model, both classical and Bayesian. We show analysis results for simulated data without repeated stimuli, as well as for resting state fMRI data. The data is analyzed both for location of the activated regions and for interactions between the activated regions.

Spatial-Temporal Modeling of Forest Gaps Generated by Colonization from Below- and Above-Ground Bark Beetle Species

Jakob G. Rasmussen, Brian Aukema, Jesper Møller, Kenneth Raffa and Jun Zhu (jgr@math.aau.dk)
Department of Mathematical Sciences, Aalborg University, Denmark

Studies of forest declines are important, because they both reduce timber production and affect successional trajectories of landscapes and ecosystems. Of particular interest is the decline of red pines which is characterized by expanding areas of dead and chlorotic trees in plantations throughout the Great Lakes Region. Here we examine the impact of two bark beetle groups, namely red turpentine beetles and pine engraver bark beetles, on tree mortality and the subsequent gap formation over time in a plantation in Wisconsin. We construct a spatial-temporal model that quantify the relations among red turpentine beetle colonization, pine engraver bark beetle colonization, and mortality of red pine trees, while accounting for correlation across space and over time. For statistical inference, we adopt a Bayesian hierarchical model and devise Markov chain Monte Carlo algorithms for obtaining the posterior distributions of model parameters as well as posterior predictive distributions.

Keywords: Autologistic model, Bayesian inference, forest entomology, Markov chain Monte Carlo, perfect simulation, spatial-temporal processes.

Directed Model Checks for Regression Models from Survival Analysis

Axel Gandy (agandy@web.de)
Centre for Advanced Study, Oslo, Norway

In survival analysis, a group of, say n , individuals is observed, that experience events over time. Let $N_i(t)$ be the number of events for the i th individual up to time t . Regression models are designed to estimate if and how certain covariates affect the counting process $N_i(t)$. Models are usually defined by the so-called intensity which is a predictable stochastic process such that $N_i(t) - \int_0^t \lambda_i(s) ds$ is a local martingale.

The best known model is the semiparametric Cox proportional hazards model, which assumes that

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{Z}_i(t)^\top \beta R_i(t)),$$

where $R_i(t)$, the so-called at-risk indicator, is an observable stochastic process taking only the values 0 and 1 which indicates whether the i th individual is at risk or not, \mathbf{Z}_i is a p -variate vector of observable processes containing the covariates. Parameters are the deterministic function λ_0 and the vector $\beta \in \mathbb{R}^p$.

Formal goodness-of-fit tests to check models are only available for a limited numbers of models.

In this talk, goodness-of-fit tests are suggested for the following large class of models, which includes many standard regression models:

$$\lambda_i(t) = f(X_i(t), \theta_v(t), \theta_c),$$

where f is a known continuous function, X_i are observable stochastic processes containing the covariates, $\theta_v(t)$ is an unknown vector-valued function and θ_c is a finite-dimensional parameter.

As a test statistic we use

$$n^{-1/2} \sum_{i=1}^n \int_0^t c_i(s) \left[dN_i(s) - f(X_i(s), \hat{\theta}_v(s), \hat{\theta}_c) ds \right],$$

where $\hat{\theta}_v$ (resp. $\hat{\theta}_c$) are estimators for θ_v (resp. θ_c) and the weights c_i are some predictable processes.

By adjusting c_i , the test can be made particularly powerful against desired alternatives, e.g. another regression model. Furthermore, one can restrict c_i such that the asymptotic distribution of the test statistic does not depend on which estimators are used for the unknown parameters.

The approach generalizes the approach of [1] and [2] to the nonlinear case. The asymptotic properties of these tests are being derived. We discuss how to choose the weights c_i suitably. If time permits, simulation studies and applications to real datasets are presented.

References:

- 1 Axel Gandy and Uwe Jensen. On goodness of fit tests for Aalen's additive risk model. *Scand. J. Statist.*, **32**:425-445, 2005.
- 2 Axel Gandy and Uwe Jensen. Checking a semi-parametric additive risk model. *Lifetime Data Anal.*, **11**:451-472, 2005.

Sparse structure of statistical dependence and feature subset selection in supervised classification

Tatjana Pavlenko (tatjana.pavlenko@miun.se)

Department of Engineering, Physics and Mathematics, Mid Sweden University, Sweden

Feature selection has been an active and fruitful field of research and development for decades in statistical classification and pattern recognition. These methods are especially important in applications since modern experimental techniques make it possible to collect and analyse huge amount of data which is mapped to a very high dimensional sample space. A typical example is micro-array experiments generating data sets with expression values for thousands of genes, but not more than a few dozen observations. A challenging task for statistical research in this area is to develop classification methods with good performance properties in such high dimensional framework.

Although the number of measured feature variables could be in the thousands, we suppose that only a few underlying features or feature subsets are strongly associated with a class variable and account for nearly all of its variation - that is, determine the class membership. This means that true underlying covariance among features assumes a sparse structure. The idea is motivated from the biological assumption that a few latent gene expression signatures (gene subsets) are most relevant for discrimination between tissue types. In this paper we focus on identification of these subsets and evaluating their ability to discriminate between classes.

We describe a supervised subset selection technique which uses distance-based score for measuring the strength of a subset of features for explaining the response variable. This technique is specially designed for classification problems in high dimension, low sample size situations. We investigate asymptotic properties of the suggested measure of separation under appropriate regularity conditions and establish its asymptotic distribution, which makes it possible to directly connect the separation score with statistical significance figures.

The performance properties and classification power of the classifier with feature selection have been studied using several simulation models and real data sets, empirically verifying the usefulness of the suggested selection technique.

On the Benjamini-Hochberg method for multiple testing

Jose A. Ferreira (j.a.ferreira@amc.uva.nl)

Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, The Netherlands

We provide a method for calculating the sample size required to attain a given average power (the ratio of rejected hypotheses to the number of false null hypotheses) and a given false discovery rate (the number of incorrect re-

jections divided by the number of rejections) in adaptive versions of the Benjamini-Hochberg method for multiple testing. The method works in an asymptotic sense as the number of hypotheses grows to infinity and under quite general conditions, and it requires data from a pilot study. The consistency of the method follows from several results in classical areas of nonparametric statistics developed in a new context of 'weak dependence'.

Wednesday, June 14

More hypotheses versus more power: Designing a multiple hypothesis testing experiment subject to a maximum overall number of possible observations

Andreas Futschik (andreas.futschik@univie.ac.at)
Dept. of Statistics, University of Vienna, Vienna, Austria

With modern complex datasets, situations involving a large number of testing problems become more and more common. In some cases, the overall number of observations m to be collected is limited, but there is some choice at the design stage concerning the allocation of the observations among the hypotheses testing problems of interest. Indeed, when m is fixed, the consideration of a large number K of hypothesis pairs implies a small average number of observations $n \simeq m/K$ for the individual problems. As a consequence, one faces small power when a multiple hypothesis testing procedure is used. Thus the overall number of rejections of null hypotheses may well be increased by restricting inference to a smaller number k of hypothesis pairs for which a larger number of observations will then be available. We are interested in finding the optimum number k to pick out of the K hypotheses. We present some general observations and provide asymptotic approximations for the optimum k as m and k both tend to infinity. We also give an illustrative example. Further details can be found in Futschik and Posch (2005).

References:

1. Futschik, A. and Posch, M. (2005) On the Optimum Number of Hypotheses when the Number of Observations is Limited. *Statistica Sinica* **15**, 841-855.

Statistical approaches in inverse problems

Jari P. Kaipio, Marko Vauhkonen, Ville Kolehmainen, and Arto Voutilainen (Jari.Kaipio@uku.fi)
Department of Physics, University of Kuopio, Finland

Bayesian model for object localization and recognition

Jouko Lampinen (Jouko.Lampinen@tkk.fi)
Helsinki University of Technology, Helsinki, Finland

Object recognition is a challenging inference task, due to internal variability of the object classes and external factors such as rotation, occlusion, and scale changes. We discuss a Bayesian model for the object matching problem, where both the localization of the objects and the learning of the object classes and their statistics correspond to computing various posterior distributions of the model given the images. For computing the posteriors we use Sequential Monte Carlo method, in which the features are matched sequentially, utilizing the information about the locations of found features to constrain the task. The non-parametric particle representation of hypotheses about the object position allow matching in multimodal and cluttered environment, where batch algorithms may have convergence difficulties. The Bayesian approach allows easy augmentation of the likelihood (appearance model) to include different distortions and variations. For example, by adding a model for occlusion of the feature point, the object can be localized from only a few visible features. We also discuss generalization of the model to include learning novel object classes from example images.

Comparing approximate methods for geostatistical inference

Jo Eidsvik, Ingelin Steinsland, and Håvard Rue (joeid@math.ntnu.no)

Department of Mathematical Sciences, NTNU, Trondheim, Norway

We study spatial data arising in geostatistical applications. The statistical model is of a hierarchical type; using a latent unobserved field, and in a Bayesian context with priors for unknown parameters.

Let $y_i, i = 1, \dots, n$, denote data at n spatial locations. Given a latent field $x_j, j = 1, \dots, m$, with $m \geq n$, the conditional likelihood of data is specified by $f(y|x, \eta)$, for some hyperparameter η . This distribution is not necessarily linear or Gaussian. The latent variable x is a Gaussian random field or a Gaussian Markov random field with covariance matrix $\Sigma(\theta)$, where θ are parameters including the spatial correlation range and the marginal variance of the field.

Our objective is to assess the marginal posteriors of hyperparameters and the latent field, i.e. $f(\eta, \theta|y)$ and $f(x|y)$. We compare two approximate methods for making such inferential statements;

- Markov chain Monte Carlo sampling from the posterior of hyperparameters and latent field.
- Numerical calculation of the marginal posteriors using Gaussian approximations for $f(x|y, \eta, \theta)$.

We also compare results based on Gaussian random fields with those using Gaussian Markov random fields as approximations.

We use data of radionuclide concentrations on Rongelap Island and seismic reflection amplitudes from a North Sea petroleum reservoir. The first dataset is acquired at irregular spatial locations, while the second is on a regular grid.

Seismic lithology-fluid prediction based on a hidden Markov random field model

Henning Omre and Marit Ulvmoen (omre@math.ntnu.no)

Department of mathematical sciences, NTNU, Trondheim, Norway

The knowledge of lithology (rock types) and fluid filling (water, oil or gas) in the reservoir is crucial in evaluation of petroleum prospects. In the North Sea these reservoirs are offshore at a depth of about 3000 meters and hence not easily assessable. The lithology-fluid characteristics must be predicted based on general knowledge about geological sedimentation and fluid behaviour, and on reservoir specific indirect observations. These observations are usually made through seismic surveys from ships and measurements in a small number of wells. The observations do not uniquely determine the lithology-fluid characteristic, hence their prediction can be considered an illposed inverse problem.

The inverse problem is cast in a Bayesian framework with a prior model representing the general knowledge about lithology-fluid properties, while the likelihood model represent the observation acquisition procedure. In the current study the lithology-fluid variables are represented by the four classes: shale, brine (water) filled sand, oil filled sand, and gas filled sand.

The prior model captures information about the vertical sequence of sedimentation of sand-shale and gravity segregation of brine-oil-gas. Moreover, the lithology-fluid characteristics are known to be fairly continuous horizontally. In order to represent this fairly precise prior knowledge, the prior Markov random field is formulated in a particular way which is coined a profile Markov random field.

The likelihood model captures information about the observation acquisition procedure. Well observations are assumed to be exact observations of lithology-fluid properties along vertical wells. The seismic data however, consists of reflections of sound pulses generated at the earths surface. These reflections are caused by changes in the lithology-fluid properties and it is modeled by wave propagation in solid matters which entails angle-dependence and convolution in the seismic data. Moreover, the observation errors are expected to have considerable spatial dependence. Consequently, the likelihood model is non-linear with strong spatial coupling.

The posterior model is fully defined by the prior and likelihood models, but the normalizing constant is not analytically tractable. Brute force MCMC sampling is hardly feasible for 3D problems of this size. We have defined

an approximate posterior model which can be assessed by simulation using a mixture of a recursive and an iterative algorithm. The algorithm appears to have favorable convergence characteristics.

The approach will be evaluated on a 2D synthetic reservoir case which is inspired by a real North Sea reservoir.

Modelling significant wave height using altimeter measurements

Anastassia Baxevasi (anastass@math.chalmers.se, baxevasi@maths.lth.se)

Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden

Significant wave height, H_s , is a measure of the variability of the ocean surface and is defined to be four times the standard deviation of the height of the ocean surface. Estimates of H_s can be considered as two dimensional random field that develops over time. Models for such spatio-temporal surfaces are generally complex, but the complexity is often constrained by the nature and limitations of the available data.

In this work, we propose a method for constructing models for estimates of H_s based on fitting random field models. The data used are that collected from the TOPEX-Poseidon satellite. We shall be considering the logarithms of the H_s values as partial observations of a random field that is well approximated by a locally stationary Gaussian random field. Moreover in a small area, the mean of the field can be modelled by means of a seasonal component and the variance is independent of time. Our interest therefore lies in modelling the residual field.

It is our experience that the residual field may be decomposed to a sum of three independent random fields that are zero-mean, isotropic with a special type of covariance structure. To estimate the parameters of the covariance directly is impossible, since the available data is along satellite tracks. Instead, we estimate the covariance parameters of the process that equals the restriction of the field along the satellite tracks, and then, under additional assumptions, we estimate the parameters in the covariance of the field. A new method is proposed using the total variation and moment estimation. For the temporal correlation of the field, the only available data having the appropriate temporal resolution is from buoys, which are traditionally located along the coast. A temporal covariance function is fitted to the data from different buoys. Finally the spatial and temporal models may be combined to give a spatio-temporal model.

Finally, the proposed model has been used to reconstruct the H_s surface on the North Atlantic, conditionally on the satellite measurements. The reconstructed surface was then compared to the surface obtained from the ERA 40 data. The agreement between the surfaces was more than satisfactory, which leads us to believe that although quite simple in its present form the model has the right variability.

A copula goodness-of-fit test based on the probability integral transform

Daniel Berg and Henrik Bakken (daniel@danielberg.no)

The Norwegian Computing Center and University of Oslo, Trondheim, Norway

Copulae have proved to be a very useful tool in the analysis of dependency structures and is now one of the main ways of modelling dependence. However, to check whether the dependency structure of a data set is appropriately modelled by a chosen family of copulae, there is no recommended method agreed upon. Prior to the use of goodness-of-fit (GOF) tests, various information criteria were employed, such as Akaike's Information Criterion (AIC). These tests do not provide us with any understanding of the size of the decision rule employed, nor its power. Hence, GOF tests are preferred.

Lately, several copula GOF tests have been proposed in literature. Panchenko (2005) propose a test based on positive definite bilinear forms, while Genest et al. (2006) propose a GOF test based on Kendall's process. Chen et al. (2004) propose two tests, both based on the probability integral transform (PIT) of Rosenblatt (1952). Their first test is consistent but suffers the curse of dimensionality. Their second test, based on the test by Breyer et al. (2003), does not have this problem. It is however inconsistent. This means that the test is not strictly increasing for every deviance from the null hypothesis, there may be deviations cancelling each other. Chen et al. (2004)'s test weights the tails of the copula, implicitly, through the squared inverse gaussian cumulative distribution function.

The PIT transforms a set of dependent variables into a set of independent variables, given the multivariate distribution. The idea of tests based on the PIT is to PIT an observed data set under a null hypothesis, and then test

for independence.

We introduce a new GOF test. The test is based on the PIT and builds on the second test by Chen et al. (2004). Our test solves the consistency problem through a transformation of the PIT data. Our test also decouples the estimation of deviance from the null hypothesis and the weighting, such that any weight function may be applied. This flexibility in the weighting function is appropriate for instance in applications where one wishes to focus more on specific areas of the copula, e.g. the tails.

Results show that our test has good power at distinguishing tail heaviness- and skewness properties. When we add tail weight, the power at distinguishing tail heaviness increases dramatically. The results also show that the second test of Chen et al. (2004) performs well for tail heaviness but very poor for skewness. Applied to a collection of stock portfolios our test strongly rejects the Gaussian and the Clayton copulae, while the Student's t copula provide a good fit.

Keywords: Copulae, Goodness-of-fit, Probability Integral Transform

References:

1. Breymann, W., A. Dias, and P. Embrechts (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance* **1**, 1-14.
2. Chen, X., Y. Fan, and A. Patton (2004). Simple tests for models of dependence between multiple financial time series, with applications to U.S. equity returns and exchange rates. Financial Markets Group, London School of Economics, Discussion Paper 483. Revised July 2004.
3. Genest, C., J.-F. Quessy, and B. Remillard (2006). Goodness-of-fit procedures for copula models based on the probability integral transform. *Scandinavian Journal of Statistics* **33**.
4. Panchenko, V. (2005). Goodness-of-fit test for copulas. *Physica A* **355**(1), 176-182.
5. Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23**, 470-472.

A kernel goodness-of-fit statistic for Pareto-type behavior

Yuri Goegebeur, Jan Beirlant and Tertius de Wet (yuri.goegebeur@stat.sdu.dk)

Department of Statistics, University of Southern Denmark, Odense, Denmark

In this paper we introduce a general kernel goodness-of-fit test statistic for assessing whether a sample is consistent with the Pareto-type model. The derivation of the class of statistics is based on the close link between the strict Pareto and the exponential distribution and puts some of the available goodness-of-fit procedures for the latter in a broader perspective. The limiting distribution for this general kernel statistic is derived under mild regularity conditions and two important special cases, corresponding to the Lewis and Jackson kernel function, will be investigated in greater depth. This approach to goodness-of-fit testing has the advantage that second order conditions on the tail behavior can be easily incorporated, resulting in bias-corrected statistics. The procedure is also computationally simple. The relation between Pareto-type goodness-of-fit testing and the optimal selection of the number of extreme order statistics for tail estimation, for instance using Hill's estimator, is examined. In this respect, we developed an algorithm based on the Lewis test statistic and compared its performance with some recently proposed procedures using a small sample simulation study. The methodology is illustrated with two practical studies.

Keywords and phrases: extreme value index, quantile-quantile plot, kernel statistic, goodness-of-fit.

Characteristics Proprieties of Long Memory Models

Karima Belaïde and Mohamed Bentarzi (email)

University a/Mira BEJAIA, ALGERIA

Log memory time series has been a topic of consideration recent interest. Much attention has been given recently to periodic processes. The aim of this paper is to studying some proprieties of a periodic long memory time series. A sufficient condition (not necessary) invertibility has been given. The expressions of the autocovariance function, autocorrelation function and spectral density are generalized to the periodic long memory processes, of a first order. In the last party, we establish the relation between periodic long memory model and periodically correlated models.

Key words and phrases: Long memory, Causality, Invertibility, Autocovariance function, Autocorrelation function, periodically correlated process.

Mathematical finance for energy markets: stochastic models and pricing of derivatives

Fred Espen Benth (fredb@math.uio.no)

Centre of Mathematics for Applications, Department of Mathematics, University of Oslo, Norway

We discuss different mean-reversion stochastic processes with jumps for modelling the evolution of temperature and spot prices of energies. The typical examples we have in mind are the weather derivatives market and the liberalized electricity and gas markets. Based on these models, we derive prices for different energy forward/futures contracts which are settled over a time period rather than at a fixed settlement time. Forward/futures contracts on temperature is usually written on some averaging over warm or cold days, and we derive explicit formulas based on a temperature model with seasonal variance. We further introduce the Heath-Jarrow-Morton (HJM) approach to energy forward/futures modelling, an approach inspired from fixed-income markets. A special emphasis is put on modelling a reasonable term-structure for the volatility of forward/futures prices. Different methods to price options are analyzed.

Some recent developments in latent variable modelling

Anders Skrondal (a.skrondal@lse.ac.uk)

Department of Statistics, London School of Economics, London, UK

Latent or unobserved variables are commonly used in modern statistical modelling. Examples include random effects in hierarchical or multilevel modelling, common factors in factor analysis, frailties in survival analysis and discrete latent variables in latent class and mixture modelling. Recently, general frameworks have been proposed that take a unified view of these different kinds of latent variable models. The frameworks can handle response variables of different types such as continuous, dichotomous, ordinal, nominal as well as counts and survival. Different types of responses can then be combined to produce for instance joint models of survival and dropout. Another useful feature of the frameworks is that latent variables may be continuous (parametric or non-parametric) or discrete or mixed continuous-discrete. Challenges include estimation and inference for complex models. The challenges become compounded when complex sampling designs are used.

Classification of *Penicillium* Fungi through Multi-Spectral Imaging and Least Angle Regression - Elastic Net Model Selection

Line H. Clemmesen, Michael E. Hansen and Bjarne K. Ersbøll (line.clemmesen@mail.dk)

Informatics and Mathematical Modelling, Technical University of Denmark, Denmark

Traditional multivariate statistical methods are adequate in situations with few variables relative to the number of observations. Unfortunately, the same methods are not applicable in most cases where the situation is reversed, i.e. there are more variables than observations. Especially, when analyzing multi-spectral images the number of variables most often exceeds the number of observations, and therefore complicates the multivariate statistical analysis.

Some problems with relative many variables in relation to observations have been solved, with success, by combining data compression techniques, e.g. Principal Components or Factor Analysis, with a subsequent method of analysis as e.g. t-tests, Discriminant Analysis etc. Furthermore, cross validation has proven to contribute in regards to variable selection [Conradsen 2002], [Skettrup 2003] and [Hastie 2001].

Newer methods have temporarily been suggested that integrates the data compression and variable selection in one step, the most recent called LARS-EN (Least Angle Regression - Elastic Net) was suggested in [Zou 2005]. The method can perform both regularization and variable selection, or one of these depending on the parameters. This makes it very useful in particular when the number of variables is much larger than the number of observations. These methods are here used for classification and for this use dummy variables representing the three classes are introduced. The newer methods are compared to the traditional.

In this study three species of the fungal genera *Penicillium* are subject to classification. *Penicillium* is one of the

most important fungal genera, as some of its species produce important drugs (e.g. penicillin) and some of its species are used in food fermentation, e.g. cheeses and mould fermented salami. All fungal species produce different mycotoxins and the methods for visual identification used are subjective [Samson & Frisvad 1993, Christensen et al. 1994], and objective identification methods are therefore highly desirable [Dörge et al. 2000, Hansen et al., 2003].

We analyze multi-spectral images of three species: *P. melanoconidium*, *P. polonicum* and *P. venetum*. The dataset consists of 36 observations, 4 isolates with 3 replications from each of the three species, from which 3754 variables are extracted from all of the 18 spectral bands in the images. It is shown that the three species of *Penicillium* can be classified correctly with both leave-one-out and 6-fold cross validation including only 6 variables by LARS-E (2 for each dummy variable).

References:

1. Christensen, M., Miller, S.L. and Tuthill, D. 1994. Color standards – a review and evaluation in relation to *Penicillium* taxonomy. *Mycol. Res.* **98**: 635-644.
2. Conradsen, L. 2002. Statistiske analyser af to-dimensionale elektroforese-geler, Master's thesis, Technical University of Denmark. (Danish)
3. Dörge, T., Frisvad, J.C. and Carstensen, J.M. 2000. Direct identification of pure *Penicillium* species using image analysis. *J. Microbiol. Meth.* **41**: 121-133.
4. Hastie, T., Tibshirani, R., & Friedman, J. 2001. The elements of statistical Learning, Springer.
5. Hansen, M. E., Lund, F. and Carstensen, J. M. "Visual clone identification of *Penicillium commune* isolates". *Journal of Microbiological Methods*, **52** (2003), pp. 221-229.
6. Samson, R.A. and Frisvad, J.C. 1993. New taxonomic approaches for identification of food-borne fungi. *Int. Biodegr. Biodet.* **32**: 99-116.
7. Skettrup, M. 2003. Multivariat dataanalyse af 2d-elektroforesegeler. Master's thesis, Technical University of Denmark. (Danish)
8. Zou, H. & Hastie T. 2005. Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B* **67**(part 2): 301-320.

Registration and shape modelling of porcine bone structures via CT

Søren G. Erbou, Rasmus Larsen, Bjarne K. Ersbøll (sge@imm.dtu.dk)

Informatics and Mathematical Modelling, Technical University of Denmark, Denmark

Based on 2D computed tomography (CT) scans of porcine carcasses, a 3D point based statistical shape model of bone structures connected to the pelvic bone, fig. 1(a), is built. The shape model is used by the Danish Meat Research Institute (DMRI) to optimize and validate the functionality of a specific tool in a slaughterhouse robot currently being developed.

The data consists of 2D CT scans, fig. 1(b), of 40 porcine carcasses separated along the medial plane. Each scan has a slice thickness of 10mm with a spacing of 10mm between each scan. Voxel dimensions are $[x, y, z] = [0.88, 0.88, 10]$ mm. The full length of the carcasses is scanned resulting in approximately 130 scans per carcass, but only 30 scans per carcass are used in this application, covering the parts around the region of the pelvic bone. Extracting corresponding points on a 3D shape from 2D scans is a tedious and difficult task, calling for (semi-)automated methods. Standard thresholding techniques combined with morphology ensures a robust segmentation of bone contours in the 2D scans. Points on the contours, fig. 1(c), are used as 3D constraints in the reconstruction of the bone surfaces using variational interpolation and radial basis functions (RBF) [3]. Due to the massive amount of data, each contour is approximated using fourier basis functions and then resampled. Only points along the contours having high curvature are selected to be constraints on the 3D surface. The implicit surface, fig. 1(d), is then resampled much more dense than the original scanning and the shapes are aligned using the iterative closest point (ICP) algorithm [1] and point correspondence is achieved. Using principal component analysis (PCA) a compact statistical shape model [2] is built describing the shape variation of the data set. A generative statistical shape allows simulation of bone structures and thereby the different conditions under which the robotic tool is to be applied. Furthermore, the generative model can be used as a prior for segmentation of new unseen pig carcass scans. The shape model is used by the DMRI to optimize the design of a tool in a new slaughterhouse robot. Applying the knowledge of how the bone structures vary, makes the process of developing new tools to do specific cuts, much less cumbersome compared to the normal trial-and-error approach.

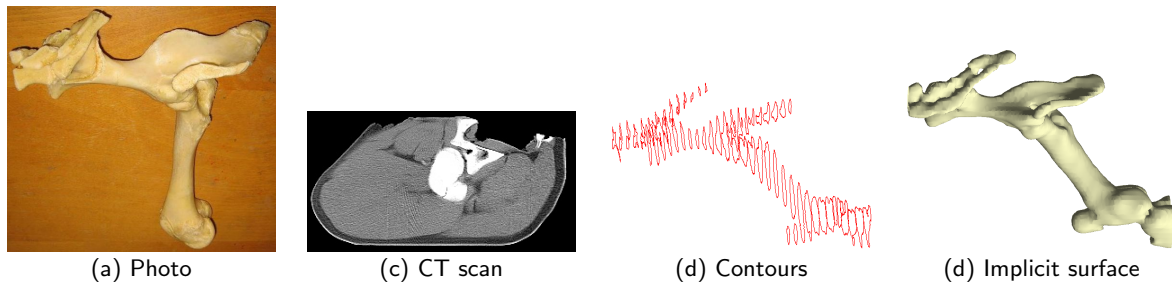


Figure 1: Pelvic bone

References:

1. Besl, P.J. and McKay, N.D. "A method for registration of 3-D shapes". Pattern Analysis and Machine Intelligence, IEEE Trans. on. Vol.14 (2), p. 239-256, 1992.
2. Cootes, T.F. and Taylor, C.J. "Statistical Models of Appearance for Computer Vision". Technical Report, University of Manchester, http://www.isbe.man.ac.uk/~bim/Models/app_models.pdf, March 2004.
3. Turk, G. and O'Brien, J. "Shape transformation using variational implicit functions". Computer Graphics Proceedings, SIGGRAPH. p. 335-342, 1999.

Predictor construction in multivariate regression: A framework and comparison

Anders Björkström (bjorks@math.su.se)

Department of Mathematics, Stockholm University, Sweden

We demonstrate that a number of well-established multivariate regression methods for prediction are related, in that they are special cases of basically one general procedure. We try a more general method based on this procedure, with two metaparameters. In a simulation study, we compare this method to ridge regression and to both multivariate PLSR and PLSR for each response separately. We find that ridge regression gives the largest errors. Among the other methods, none is in any obvious way superior to the others, although the new method runs a somewhat larger risk to perform very poorly. We also illustrate how method performance is affected by the degree of near-collinearity in the X data, and, to a smaller extent, in the (true) coefficients matrix.

Small sample bias and selection bias effects in calibration under latent factor regression models

Rolf Sundberg (rolfs@math.su.se)

Mathematical statistics, Stockholm University, Sweden

We study bias of predictors when a multivariate calibration procedure has been applied to relate a scalar y (concentration of an analyte, say) to a vector x (spectral intensities, say). The model for data is assumed to be of latent factor regression type, with multiple regression models and errors-in-variables models as special cases. The calibration procedures explicitly studied are OLSR, PLSR and PCR. When y has been systematically selected in the calibration to achieve increased variation, which is often the case in practice, this leads to biased coefficients in the predictor, in particular seen when validation set y -values are regressed on predicted y (a selection effect). Another bias effect is a sample size effect, increasing with reduced calibration sample size and with increasing dimension of x (absent when x is univariate). Formulae are given for these bias effects, both separately and in combination, and the formulae are illustrated and compared with simulation model results. As a qualitative example, PLSR and PCR are less sensitive to small samples than OLSR, but equally sensitive as OLSR to selection.

Analysis of Classical Swine Fever Virus Incidence Data using a Partial Likelihood Approach

Michael Höehle and Inga Tschöpe (hoehle@stat.uni-muenchen.de)

Department of Statistics, University of Munich, Germany

In this work we analyse data provided by the Federal Research Centre for Virus Diseases of Animals, Wusterhausen, Germany, on the incidence of classical swine fever virus (CSFV) in Germany during 1993-2004. Of interest is the spatio-temporal connection between the incidence among wild boars and domestic pigs. To this end, the data contain information on the individual pig farms infected by CSFV from four German federal states, together with the incidence rates among wild boars established from the analysis of hunting bags. A partial likelihood approach for spatio-temporal point processes proposed by Diggle (2005), used to model data from the 2001 UK foot-and-mouth epidemic, is adapted to the CSFV setting of coarser spatial resolution and multiple outbreaks. The idea of the Diggle (2005) approach is the following: by conditioning on the locations and time-points of the events, a log-likelihood for the observed time order of the events can be formulated. Similar to Cox's proportional hazards model, the conditional intensities contain a base hazard function, which is then considered nuisance. Proportional to this base hazard the intensity, by which a susceptible farm is infected by an infectious farm, is modelled as the product of a distance kernel and (possibly time-varying) covariate effects of the susceptible and infectious farm, respectively. Inference based on this partial likelihood is now straightforward and can be done using standard maximization algorithms such as Nelder-Mead or BFGS. We fit the above model to the CSFV data using the `optim()` routine in R, report profilelikelihood based standard errors and use Akaike's information criterion to decide on the inclusion of covariates and various distance kernel representations. A connection between wild boar and domestic pig CSFV-incidence is found, but the spatial resolution of the data only allowed us to substantiate a very limited neighbourhood interaction between farms.

Space-time point processes applied to the modeling of fire occurrences

Carlos Diaz (carlos@sigma.iimas.unam.mx)

IIMAS, UNAM, Mexico

In North America, it is well known that fire is a crucial component of forest ecology. Therefore, it is necessary to model forest fire occurrences, either to plan fire control activities or as a base for landscape change models among other applications. Because fire ignitions may be considered as point events in principle, an appealing way for modeling fire occurrences is by using point process models.

In this paper we present an application of space time modeling to fire occurrences in NW North America using a point process approach. We model the intensity function as dependent on covariates related to fire risk factors. The intensity function also depends on fire occurrences in nearby occurrences in space-time, in order to incorporate the recovery time of burned areas. Simulation-based methods were used for parameter estimation, as well as maximum likelihood.

The results provide estimates of high risk areas, as well as insight on the effect of the risk factors considered in the analysis.

Perfect simulation for posterior mixture weights

Kasper Klitgaard Bertelsen (k.berthelsen@lancaster.ac.uk)

Lancaster University, UK

We consider a standard Bayesian analysis of the unknown mixture weights for a mixture of known densities. It is well known how to approximately sample the posterior weights by Gibbs sampling using the duality principle, i.e. introducing auxiliary allocation variables.

In this talk we combine a number of recent techniques to formulate an algorithm for perfect simulation of the posterior weights. The algorithm is an example of Wilson's read-once coupling from the past algorithm (read-once CFTP). The key ingredient of this algorithm is constructing compounds of random maps so that stationarity is preserved and there is a positive probability that the compound map is coalescent, i.e. maps the state space into a single point.

Finding a random map that preserves stationarity is generally easy. In contrast determining if the compound map is coalescent is in general hard. This is made much easier if the state space has a partial ordering and the random map is (anti-)monotone w.r.t. this ordering. Here, having more than two components this is not the case.

The Gibbs sampler can be represented as a random map. We show that it is possible to find upper and lower bounds on the image of the resulting random map which suffice to formulate an algorithm for perfect simulation. In cases where the mixture component densities are close, the bounds we obtain are too "sloppy" making detection of coalescence difficult and consequently perfect simulation inefficient.

To alleviate this problem we consider catalytic updates as introduced by Breyer and Roberts. A catalytic update is a random modification of an existing random map. The resulting modified map maps (a subset of) the state space into a single point. Applying enough catalytic updates it is hoped that the state space is mapped into a (small) finite number of states. Coalescence is then easily checked by tracking the finite states.

The problem is that applying catalyst and checking if they absorb the entire state space is computationally very expensive. In the spirit of one-shot coupling we suggest to delay the application of the catalytic update until it is advantageous.

Our method can be extended to the case where the mixture densities are specified by unknown parameters. In practice our approach is of limited use in this setup as exact sampling is feasible only for small data sets.

Finally we show how our approach can be extended to hidden Markov models. Specifically we consider a two state hidden Markov model and generate exact samples from the posterior transition probabilities. In this case we only rely on upper and lower bounds as the lack of conditional independence of the data makes catalytic updates infeasible.

This is based on joint work with Laird A. Breyer and Gareth O. Roberts.

Control variates for the Metropolis-Hastings algorithm

Hugo Hammer and Håkon Tjelmeland (hammer@math.ntnu.no)

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Consider a Metropolis-Hastings algorithm for simulating realisations from a given target distribution $\pi(x)$, $x \in \mathbb{R}$. Letting x denote the current state of the Markov chain, each iteration consists of first proposing a potential new state y from some proposal density $q(y|x)$ and thereafter accepting y as the new state with an acceptance probability $\alpha(y|x)$, otherwise keeping x as the current state. After discarding a burn-in period, the current states are essentially from the target distribution $\pi(\cdot)$ and can be used to estimate mean values $\mu = E[f(x)] = \int f(x)\pi(x) dx$. A natural and the most frequently used estimator of μ is simply the empirical mean of $f(x)$.

In this talk we consider the use of control variates to improve the empirical mean estimator. Then μ is estimated by a linear combination of the original empirical mean estimator and some control variates. Examples of control variates used in a Metropolis-Hastings setting can be found in Pinto and Neal (2001), Mira et al. (2003) and Atchadé and Perron (2005). Our key idea in this talk, see also Hammer and Tjelmeland (2005), is to use control variates that are functions of both the current state, x , and the proposal, y . The simplest alternative is then to define the control variate as

$$g(x, y) = w_1(x, y)f(x) + w_2(x, y)f(y),$$

where $w_1(x, y)$ and $w_2(x, y)$ are weight functions. Choosing

$$w_1(x, y) = -w_2(x, y) = \frac{\pi(y)q(x|y)}{\pi(x)q(y|x) + \pi(y)q(x|y)}$$

it easily follows that $E[g(x, y)] \equiv \int \int g(x, y)\pi(x)q(y|x) dx dy = 0$ for general target and proposal distributions. We also define a number of zero mean control variates that are functions of the acceptance indicator of the Metropolis-Hastings algorithm. We present empirical results for four simulation examples, also including a reversible jump situation. The variance reduction varies depending on the target distribution and the proposal mechanisms used. The simplest control variate (defined above) generally seems to give the best results. In one example we get a variance reduction of 45%, but more typical reductions are between 15% and 35%.

References:

1. Atchadé, Y. F. and Perron, F. (2005). Improving on the independent Metropolis-Hastings algorithm, *Statistica Sinica* **15**: 3–18.
2. Hammer, H. and Tjelmeland, H. (2005). Control variates for the Metropolis-Hastings algorithm, Technical report, Statistics no. 8, Department of Mathematical Sciences, Norwegian University of Science and

Technology, Trondheim, Norway.

3. Mira, A., Tenconi, P. and Bressanini, D. (2003). Variance reduction in MCMC, Technical Report 29/2003, Department of Economics, University of Insubria, Italy.
4. Pinto, R. L. and Neal, R. M. (2001). Improving Markov chain Monte Carlo estimators by coupling to an approximating chain, Technical Report No. 0101, Department of Statistics, University of Toronto.

Using RjMCMC Algorithm to Estimate Supply and Demand Curves

Elena Moltchanova, M. Obersteiner and S. De Cara (email)

International Institute for Applied System Analysis, Laxenburg, Austria

The aim of this study was to model the global supply and demand curves for wood products during the last decade. The data on import and export quantities and value and therefore on the prices was obtained from Food and Agriculture Organization (FAO) of the United Nations. The demand and supply functions for a product i were assumed to be linear of the form

$$Y_i = \sum_j \epsilon_{ij} P_j + \gamma * GDP$$

where Y_i is the demanded or supplied quantity of the good i , ϵ_{ij} is the price elasticity of good i with respect to good j .

Bayesian methods were applied to the analysis of the above data. While the global model - a single regression for all countries - was considered to be too general, the sparse nature of the data (1-10 observations per country) would cause wide posterior distributions. Therefore it was suggested to group the countries with similar demand/supply functions and to estimate the parameters for those together. The number of groups was assumed to be unknown and the grouping itself, i.e. the allocation of countries between different groups, was also to be estimated. The model therefore was rewritten in the form:

$$Y_{ci} | \mu_{ci}, \tau_{ci}, z_c \sim \mathcal{N}(\mu_{ci}, \tau_{ci}^{-1}) \quad \forall c = 1, \dots, C, \quad \forall i = 1, \dots, I$$

$$\mu_{zi} | \epsilon_{zi}, P_c, \gamma_z, GDP_c = \sum_j \epsilon_{zij} P_c j + \gamma_{z_c} * GDP_c$$

$$p(z_c = m | w.) = \omega_m \forall m = 1, \dots, k$$

where z_c is the group to which a country c belongs and the number of groups is denoted by k and is itself an object of our inference. Parameters $\{\epsilon\}$, $\{\gamma\}$, $\{\omega\}$ and k were assigned non-informative priors.

A Reversible jump - Markov chain Monte Carlo (RjMCMC) was applied first to the simulated and then to the actual data with encouraging results. Further analysis on a better (more complete/extensive data set) would be of interest.

Bayesian CART - prior specification and posterior simulation

Yuhong Wu, Håkon Tjelmeland and Mike West (Haakon.Tjelmeland@stat.ntnu.no)

Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Search methods for CART (classification and regression tree) models are traditionally based on greedy algorithms that generate trees by recursive partitioning of sample data into more and more homogeneous subsets, often followed by pruning to address over-fitting (Breiman et al., 1984; Clark and Pregibon, 1992). In contrast, Bayesian Monte Carlo methods, introduced by Chipman et al. (1998) and Denison et al. (1998), aim to generate samples of tree models according to their posterior probabilities representing fit to the data.

We present advances in Bayesian modeling and computation for CART models. The modeling innovations include a formal prior distributional structure for tree generation - the *pinball prior* - that allows for the combination of an explicit specification of a distribution for both the *tree size* and the *tree shape*. The core computational innovations involve a novel Metropolis-Hastings method that can dramatically improve the convergence and mixing properties of MCMC methods of Bayesian CART analysis. Earlier MCMC methods have simulated Bayesian CART models using very local MCMC moves, proposing only small changes to a "current" CART model. Our new Metropolis-Hastings move makes large changes in the CART tree, but is at the same time local in that it leaves unchanged the partition of observations into terminal nodes. We evaluate the effectiveness of the proposed algorithm in two

examples, one with a constructed data set and one concerning analysis of a published breast cancer data set.

References:

1. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees, The Wadsworth statistics/probability series, Belmont CA.
2. Chipman, H., George, E. and McCulloch, R. (1998). Bayesian CART model search (with discussion), J. Am. Statist. Ass. **93**: 935-960.
3. Clark, L. and Pregibon, D. (1992). Tree-based models, in J. Chambers and T. Hastie (eds), Statistical Models in S, Wadsworth & Brooks/Cole computer science series, Wadsworth & Brooks/Cole Advanced Books & Software.
4. Denison, D., Mallick, B. and Smith, A. (1998). A Bayesian CART algorithm, Biometrika **85**: 363-377.

Thursday, June 15

A martingale random effects model for longitudinal data with dropout

Robin Henderson (Robin.Henderson@ncl.ac.uk)

School of Mathematics and Statistics, University of Newcastle upon Tyne, UK

The problem of analysing longitudinal data complicated by possibly informative dropout has had considerable attention in the statistical literature. Most authors have concentrated on either methodology or application but we begin by arguing that more attention could be given to study objectives and the relevant targets for inference. Next we suggest a new and computationally efficient modelling and analysis procedure. We assume a dynamic linear model for the expected increments of a constructed variable, from which valid inference for the underlying dropout-free population is possible provided one important assumption holds. This is that subject-specific random effects follow a martingale process in the absence of dropout. An informal diagnostic procedure to assess the tenability of the assumption is proposed.

A different perspective on inverse probability weighting for causal inference

Vanessa Didelez (vanessa@stats.ucl.ac.uk)

Department of Statistical Science, University College London, UK

Inverse probability weighting is a common method when dealing with missing values. It is also useful for causal inference, which can be regarded as a missing data problem because the outcome of an individual had he/she received a different treatment from the one he/she actually received is missing. I will show how inverse probability weighting for causal inference can be motivated from a different point of view, based on a graphical approach and a notion of causality based on interventions. The cases of survival outcomes with time dependent confounding as well as sequential treatments will be paid particular attention.

Semiparametric additive hazards model with focus on a change-point model

Torben Martinussen (torbenm@dina.kvl.dk)

Department of Mathematics and Physics, The Royal Veterinary and Agricultural University, Denmark

The Aalen additive hazards model (Aalen, 1980) is a useful alternative to the Cox model when analysing survival data. A particular useful aspect of the Aalen model is that it allows for timevarying covariate effects. It will often be of interest to investigate whether a covariate is constant with time, however. In this talk we describe how such hypothesis may be investigated and how do it in practice using the R-package `timereg`. An interesting model in between the constant effects model and the full Aalen model is the change-point model where the effect of a given covariate is constant up to an unknown point in time and changed thereafter to a new value. For instance, a given new drug may have an initial effect only; or the other way around, there may an initial phase without effect. We will also describe how to do estimation and inference within such a changepoint model. The methods will be illustrated on real datasets.

Model checking techniques for grouped survival data models

Christian Bressen Pipper (c.pipper@biostat.ku.dk)

Department of Biostatistics, University of Copenhagen, Denmark

Epidemiology research often entails the analysis of failure times subject to grouping. In large cohorts interval grouping also offers a feasible choice of data reduction to actually facilitate an analysis of the data. Based on an underlying Cox proportional hazards model for the exact failure times one may deduce a grouped data version of this model which may then be used to analyse the data. The model bears a lot of resemblance to a generalised linear model, yet due to the nature of data one also needs to incorporate censoring. In the case of non ignorable censoring this precludes model checking procedures based on ordinary residuals as calculation of these requires knowledge of the censoring distribution. In this talk we represent interval grouped data in a dynamical way using a counting process approach. This enables us to identify martingale residuals which can be computed without knowledge of the censoring distribution. We use these residuals to construct graphical as well as numerical model checking procedures. An example from epidemiology is provided.

Key words: Cox regression, goodness of fit, grouped data models, interval-grouped failure times, martingale residuals, multiplier processes.

Discriminative Estimation via an Exponential Large Deviation Property

Niels Richard Hansen (richard@math.ku.dk)

Department of Applied Mathematics and Statistics, University of Copenhagen, Denmark

Several problems in biological sequence analysis, such as local alignment, can be viewed as extended change-point problems in the sense that one tries to detect and localise within a large number of random variables some that show a distributional deviation. We focus on a resulting integer programming problem where the statistic for detecting distributional deviations is given as the maximum over a random integer polytope for a given parameter, or weight, vector v . The actual optimisation is in examples of practical interest carried out by a dynamic programming algorithm. Thus we accept to work with a specific, model independent class of statistics parameterised by v – mostly justified in practice by their computational virtues. In current applications, the choice of parameter v is often somewhat ad hoc, like in gapped alignment where matches/mismatches are often scored using a log-likelihood ratio under a no-gap model (PAM/BLOSUM), and gap scores have little theoretical justification, see e.g. *Altschul et al. (1996, Meth. Enzymol. 266, 460-480)*. We deal with choosing the parameter vector v with the objective of discriminating most efficiently between a null model and a (local) alternative. That is, we seek a choice of parameters such that we obtain a large maximum when a (local) alternative is present in the data while retaining control over the maximal value under the null model.

The abstract setup we will investigate is as follows. We have a finite (in practice, huge) *configuration* space \mathcal{A} , a random map $\Phi : \mathcal{A} \rightarrow \mathbb{N}_0^d$, and a d -dimensional vector space L of parameters. Then for each $v \in L$ we assign the score $v^t \Phi(\alpha)$ to the configuration α . We seek an understanding of the distribution of the random polytopes spanned by $\Phi(\alpha), \alpha \in \mathcal{A}$ in \mathbb{R}^d under the null model as well as under the alternative. We propose considering an asymptotic framework and an exponential large deviation assumption on the distribution of $\max_{\alpha \in \mathcal{A}} v^t \Phi(\alpha)$ (for $\alpha \in C \subset L$) that provides a normalisation of the optimal score under the null model. The assumption holds for e.g. gapless local alignment, *Dembo et al. (1994, Ann. Prob. 22, 2022-2039)*, and is, furthermore, widely believed to hold for more general problems like local alignment with affine gap penalties. We propose the following two-step empirical procedure. First, from a dataset under the null model, we estimate as a function of v the normalisation parameters using a peaks-over-a-threshold (POT) procedure. The vector space L is divided into cones and on each cone the resulting normalisation becomes a rational function – actually a fraction of affine functions. The cones

form a polyhedral complex in L called the *normal fan of a Newton polytope*, which can be computed using the *polytope propagation algorithm*, Pachter et al. (2004, PNAS, 101(46), 16138-16143). Second, from a dataset under the alternative, we investigate the empirical distribution of the (empirically) normalised maximal score as a function of v , and we seek to 'maximise the location of the distribution'. We consider maximising the expectation and maximising a quantile. To investigate the properties of the proposed procedure we give a detailed theoretical analysis of a simple random walk model where, for instance, the normalisation has a rather explicit representation and where the computational problems are easier to deal with. We suggest, however, that the approach can be applied to much more advanced problems of real, biological interest, and that it provides a rational way of estimating the full parameter vector v with the objective of discrimination.

Key words: Biological sequence analysis, discrimination, fractional programming, local alignment, large deviations, normal fans, parametric inference, polytopes, polytope propagation, POT, structural prediction.

Modelling of CGH-data with continuous-index hidden Markov models

Susann Stjernqvist (susann@maths.lth.se)

Centre for Mathematica Sciences, Lund University, Sweden

Normally humans have two copies of most of their DNA, one from their mother and one from their father. Sometimes though, aberrations arise in the number of copies of segments of the DNA sequence. One way to study these alterations is array Comparative Genomi Hybridisation (array CGH). With this technique sample DNA and reference DNA are labelled with two different dyes, and then the mixture of those is hybridised on a micro array. The micro array is produced by clones, i.e. short segments of the DNA sequence which start and stop at predefined positions. A laser is used to fluorescent the dyes, and by comparing the intensities of the two colours one yields the ratio of the number of copies of the sample DNA and the reference DNA. The ratios are translated into \log_2 -scale, and since there are measurement errors they are observed in noise.

Friedland et. al. [1] used hidden Markov models to model the copy numbers of DNA. Such a model assigns to each clone an unobserved state variable, that represents the copy number, and one observed value that is the \log_2 -ratio disturbed by noise. The state variables form a finite-state Markov chain, and equal noise variances are assumed for all states. One feature of the data used in our study is that the clones overlap, i.e. one clone may start before the previous has ended, and another one is that the clones differ in length. These characteristics make a discrete-index model less suitable than a continuous-index model. We thus use a hidden Markov model with continuous index, i.e. a model in which the Markov chain can jump at any base-pair position and not only between clones. The parameters used to describe the model are the \log_2 -ratios, $\mu = (\mu_1, \dots, \mu_m)$, where m is the number of states, the measurement error variance σ_2 , which we assume is equal for all states, and the jump intensities, $Q = \{q_{ij} : i, j = 1, \dots, m\}$ of the Markov chain.

The model parameters are estimated with a Monte Carlo EM algorithm, which is a modification of the EM algorithm where the E-step is approximated with several MCMC realisations of the Markov chain. Since the number of jumps of the Markov chain changes, we use a reversible jump MCMC algorithm due to Ball et. al. [2] to generate the realisations.

One advantage of the continuous-index model is that we can estimate the exact position of the jump from one state to another, without depending on the clone positions. Another advantage is that the correlation between the residuals of adjacent clones tends to decrease compared to the residuals from the model with discrete index.

References:

1. Hidden Markov models approach to the analysis of array CGH data, Journal of Multivariate Analysis **90**, 132 – 153, 2004
2. Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using Markov chain Monte Carlo, Proceedings of the Royal Society **455**, 2879 – 2932, 1999.

Statistical Testing Within the Gene Ontology Hierarchy

Clara-Cecilie Günther, **Mette Langaas**, Stian Lydersen, Vidar Beisvåg, Frode K. R. Jünge, Hallgeir Bergum, Astrid Lægreid (Mette.Langaas@math.ntnu.no)

Department of Mathematical Sciences, The Norwegian University of Science and Technology, Trondheim, Norway

The overall aim of Systems Biology is to come to an understanding of complete biological systems. Different sources of data may enter into the modelling of the systems, e.g. microarray experiments for measuring gene expression. A popular aim of gene expression experiments is to arrive at one (or several) set(s) of reporters (genes, probe sets, ESTs) that are found to be differentially expressed between two situations (e.g. treatment A vs. treatment B). To aid in the interpretive challenge of summarizing the findings present in the obtained lists of differentially expressed genes, a strategy called gene-class testing (GCT) has been proposed. Gene classes may be based on Gene Ontology (GO) categories.

eGOn (explore Gene Ontology), <http://www.genetools.no>, is a GO-tool where lists of reporters can be submitted through a web interface to an annotation database, and automatically translated into GO terms annotated to these reporters. In addition to powerful graphical displays eGOn offers statistical hypothesis testing to assess the level of similarity between two reporter lists.

Let us consider two reporter lists; list A and list B. At the given GO-node G , we are interested in testing whether the probability of belonging to GO-node G is different for reporter list A and reporter list B. I.e. for each reporter on list A, there is a probability $P(G|A)$ of belonging to GO-node G , and for each reporter on list B, there is a probability $P(G|B)$ of belonging to GO-node G . Under the null hypothesis these two probabilities are equal.

Due to possible dependencies between the lists A and B, we distinguish between three situations:

1. One of the two reporter-lists compared is the list containing all reporters present (i.e. available on the microarray slide) at that GO-node. Conditionally, this results in a hypergeometric situation, and Fisher's exact test may be used (implemented in eGOn).
2. When the two reporter lists, A and B, are mutually exclusive there are no reporters that are on both lists, e.g. A is a list of reporters associated with up-regulation and B is a list of reporters associated with down-regulation. This results in testing two independent binomial probabilities, and Fisher's exact test may be used (implemented in eGOn).
3. Finally, the two reporter lists compared, may contain reporters that are on both lists, e.g. A is a list of reporters associated with treatment A and B is a list of reporters associated with treatment B.

For most GO-tools only the situation (a) above is handled.

The focus of this presentation is on studying and developing different statistical hypothesis tests for handling the situation (c) above. The tests are compared and evaluated in a simulation study.

General linear models for microarray experiments

Anders Sjögren, Erik Kristiansson, Olle Nerman and Mats Rudemo

(anders.sjogren@math.chalmers.se)

Mathematical Statistics Chalmers University of Technology, Gothenburg, Sweden

The nature of data from microarray experiments brings interesting statistical challenges. Typically, the expression of thousands of genes is measured simultaneously in a relatively small number of biological samples and one test is performed for each gene, e.g. to identify genes with different expression in different groups. An important challenge is then to use the structure of the data to improve the performance of the individual tests.

Previously, the vast number of measured genes have been utilised to identify a prior distribution for the variance of each gene, in effect moderating the individual variance estimates. In [1,2] we complement this by suggesting that data from different arrays may have different variability, e.g. due to random deviations in quality from some of the measurement steps. Furthermore, data from arrays may be correlated, e.g. due to similarities in the quality deviations. The result is a generalised linear model with common covariance structure for the different genes and a prior distribution for gene-wise variance scales, giving rise to weighted moderated F-tests. Estimation of the covariance structure is feasible under certain conditions and the estimate is then precise thanks to the large number of genes.

Results on real data indicate that correlations often exist between arrays and that inference results from procedures not taking this into account may be biased. In particular, the empirical distribution of p-values is sometimes drastically changed when the covariance structure is introduced in the model.

References:

1. Erik Kristiansson, Anders Sjögren, Mats Rudemo, and Olle Nerman (2005) Weighted Analysis of Paired Microarray Experiments, *Statistical Applications in Genetics and Molecular Biology*: 4(1), Article 30.
2. Erik Kristiansson, Anders Sjögren, Mats Rudemo, and Olle Nerman (2006) Quality Influenced Analysis of General Paired Microarray Experiments, Chalmers University of Technology; Department of Mathematical Sciences Preprint 2006:6, <http://www.math.chalmers.se/Math/Research/Preprints/2006/6.pdf>

Statistical inference in population genetics

Rasmus Nielsen (rasmus@binf.ku.dk)

Centre for Bioinformatics, University of Copenhagen, Denmark

Population genetics is the study of genetic variability within and between populations. It has recently gained new importance as population genetic methods are used in many areas of medical research, particularly in association mapping studies. However, inference in population genetic models present special computational and statistical problems. Most models are based on well-described stochastic process theory, but likelihood functions can usually only be calculated using computationally slow Monte Carlo procedures. The fundamental problem is that samples from different individuals are not iid but are correlated through an underlying genealogical tree. Even in some of the most simple models the likelihood function cannot be calculated without the use of very slow MCMC or Sequential Importance Sampling methods - and for many problems even these methods do not work for realistic sized samples. Researchers must often instead rely on Ad hoc inference procedures with poor or unknown statistical properties. In this talk I will give an introduction to some of the statistical problems encountered in population genetics and discuss some of the solutions that have been proposed.

Multidimensional local false discovery rate

Yudi Pawitan (yudi.pawitan@ki.se)

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

The false discovery rate (fdr) is a key tool for statistical assessment of differential expression (DE) in microarray studies. It is, however, well known that overall control of the fdr alone is not sufficient to address the problem of genes with small variance, which suffer from a disproportional high rate of false positives. Graphical tools and modified test statistics have been proposed for dealing with this problem, but there is currently no procedure for controlling the fdr directly. Methods: We generalize the local fdr called fdr2d - as a function of multiple statistics, combining a common test statistic for assessing differential expression with standard error information. Results: The fdr2d allows an objective assessment of differential expression as a function of gene variability. Furthermore, the fdr2d has comparable performance to other methods that model the variance explicitly or to the theoretically optimal procedure.

List of participants

Name (email)	Affiliation
Alexander Sokol (arp@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Alexandra Jauhiainen (alexandra.jauhiainen@math.chalmers.se)	Mathematical Statistics Chalmers University of Technology, Sweden
Amusa Sulaimon (kwarapolytechnic@yahoo.com)	Kwara Polytechnic Lagos Island, Nigeria
Ana Garcia Lopez (aglo@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Anastassia Baxevani (anastass@math.chalmers.se)	Chalmers University of Technology Gothenburg, Sweden
Anders Björkström (bjorks@math.su.se)	Mathematical Statistics Stockholm University, Stockholm, Sweden
Anders Rahbek (rahbek@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Anders Rønn Nielsen (arp@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Anders Sjögren (anders.sjogren@math.chalmers.se)	Mathematical Statistics Chalmers University of Technology, Göteborg, Sweden
Anders Skrondal (a.skrondal@lse.ac.uk)	Department of Statistics London School of Economics, London, UK
Anders Tolver Jensen (tolver@kvl.dk)	Department of Natural Sciences The Royal Veterinary and Agricultural University Copenhagen, Denmark
Andreas Futschik (andreas.futschik@univie.ac.at)	Department of Statistics University of Vienna, Austria
Antti Penttinen (penttine@maths.jyu.fi)	Department of Mathematics and Statistics University of Jyväskylä, Finland
Asger Roer Pedersen (arp@dmu.dk)	Danmarks Miljøundersøgelser Silkeborg, Denmark
Astrid Lunde (Astrid.Lunde@mfr.uib.no)	Department of Mathematics University of Bergen, Bergen, Norway
Axel Gandy (agandy@web.de)	Centre for Advanced Study Oslo, Norway
Bent Jørgensen (bentj@stat.sdu.dk)	Forskningsenheden for Statistik University of Southern Denmark, Odense, Denmark
Birgitte Biilmann Rønn (bbr@novonordisk.com)	Novo Nordisk A/S Bagsværd, Denmark
Bjarne Kjær Ersbøll (be@imm.dtu.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Denmark
Bo Markussen (bomar@kvl.dk)	Department of Natural Sciences The Royal Veterinary and Agricultural University Copenhagen, Denmark
Bo Martin Bibby (bibby@biostat.au.dk)	Department of Biostatistics Institute of Public Health, University of Aarhus, Denmark
Carlos Diaz-Avalos (carlos@sigma.iimas.unam.mx)	IIMAS, National University of Mexico Mexico City, Mexico

Name (email)	Affiliation
Christian Phipper (c.pipper@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Claus Dethlefsen (aas.claus.dethlefsen@nja.dk)	Aalborg Hospital Aarhus University Hospital, Aalborg, Denmark
Claus Holst (ch@ipm.hosp.dk)	Institute of Preventive Medicine Copenhagen University Hospital, Copenhagen
Daniel Berg (daniel@danielberg.no)	Norwegian Computing Center University of Oslo, Trondheim, Norway
David Spiegelhalter (david.spiegelhalter@mrc-bsu.cam.ac.uk)	MRC Biostatistics Unit Cambridge, UK
De Serio Clelia (diserio.clelia@hsr.it)	Vita-salute san Raffaele University Milan, Italy
Elena Moltchanova (elena.moltchanova@ktl.fi)	IIASA, KTL Helsinki, Finland
Eli Vibeke Olsen (evo@danishmeat.dk)	Danish Meat Research Institute Roskilde, Denmark
Emmanuel Abatih (ena@kvl.dk)	The Royal Veterinary and Agricultural University Denmark
Erik Kristiansson (erikkr@math.chalmers.se)	Mathematical Statistics Chalmers University of Technology, Göteborg, Sweden
Erik Lindström (erikl@maths.lth.se)	Centre for Mathematical Sciences Lund, Sweden
Erik Parner (parner@biostat.au.dk)	University of Aarhus Denmark
Ernst Hansen (erhansen@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Eva B. Vedel Jensen (eva@imf.au.dk)	Department of Mathematical Sciences Aarhus University, Denmark
Flemming Skjøth (fls@landscentret.dk)	Danish Cattle Federation Århus, Denmark
Fred Espen Benth (fredb@math.uio.no)	Centre of Mathematics for Applications University of Oslo, Norway
Gorm Gabrielsen (gg.mes@cbs.dk)	Center for Statistics Copenhagen Business School, Denmark
Gunnar Hellmund (hellmund@imf.au.dk)	Institut for Matematiske Fag Aarhus Universitet, Denmark
Gunnar Rosenqvist (gunnar.rosenqvist@hanken.fi)	Swedish School of Economics Helsinki, Finland
Hanne Wist Rognebakke (hanne.rognebakke@nr.no)	Norwegian Computing Center Oslo, Norway
Hege Marie Bøvelstad (hegembo@math.uio.no)	Department of Mathematics University of Oslo, Norway
Heino Bohn Nielsen (Heino.Bohn.Nielsen@econ.ku.dk)	Department of Economics University of Copenhagen, Denmark
Helle Rootzen (hero@imm.dtu.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Denmark
Helle Sørensen (helle@dina.kvl.dk)	Department of Natural Sciences The Royal Veterinary and Agricultural University Copenhagen, Denmark

Name (email)	Affiliation
Henning Omre (omre@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Henrik Madsen (hm@imm.dtu.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Denmark
Holger Rootzén (rootzen@math.chalmers.se)	Department of Mathematical Statistics Chalmers, Göteborg, Sweden
Hugo Lewi Hammer (hammer@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Håkon Tjelmeland (Haakon.Tjelmeland@stat.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Inge Bjørn Myrseth (inge.myrseth@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Inge Henningsen (inge@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Ivar Heuch (Ivar.Heuch@uib.no)	Department of Mathematics University of Bergen, Norway
Jakob Gulddahl Rasmussen (jgr@math.aau.dk)	Department of Mathematical Sciences Aalborg University, Denmark
Janeli Sarv (sarv@math.chalmers.se)	Mathematical Statistics Chalmers University of Technology, Göteborg, Sweden
Jari Kaipio (jari.kaipio@uku.fi)	University of Kuopio Finland
Jens Ledet Jensen (jlj@imf.au.dk)	Department of Mathematical Sciences University of Aarhus, Denmark
Jesper Møller (jm@math.aau.dk)	Department of Mathematical Sciences Aalborg University, Denmark
Jo Eidsvik (joeid@math.ntnu.no)	Department of Mathematical Sciences NTNU, Trondheim, Norway
Jon Michael Gran (j.m.gran@medisin.uio.no)	Department of Biostatistics University of Oslo, Norway
José António Ferreira (j.a.ferreira@amc.uva.nl)	Dept. of Clinical Epidemiology and Biostatistics AMC, University of Amsterdam, The Netherlands
Jouko Lampinen (Jouko.Lampinen@tkk.fi)	Helsinki University of Technology Espoo, Finland
Judith L. Jacobsen (jlj_karlebo@yahoo.dk)	Statcon ApS Kokkedal, Denmark
Jurate Saltyte Benth (jurate@ahus.no)	Helse Øst Health Services Research Centre Lørenskog, Norway
Jørgen Holm Petersen (J.H.Petersen@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Kadima Kalala Bay Mazele (mazelebay@yahoo.fr)	Ecole Pigier Fes Morocco
Kamille Sofie Tgholt (kamille@stud.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Karima Belaïde (k_tim2002@yahoo.fr)	Department of Mathematics, Universita A/Mira Bejaia, Algeria

Name (email)	Affiliation
Kasper Klitgaard Berthelsen (k.berthelsen@lancaster.ac.uk)	Department of Mathematics and Statistics Lancaster University, United Kingdom
Kasper Kristensen (kkr@dfu.min.dk)	DFU Copenhagen, Denmark
Kim Emil Andersen (kiean@vestas.com)	Vestas Asia Pacific A/S Randers, Denmark
Lasse Heikkinen (lasse.heikkinen@uku.fi)	Department of Physics University of Kuopio, Finland
Leslie Foldager (Lfo@psykiatri.aaa.dk)	Centre for Basic Psychiatric Research Aarhus, Denmark
Liberato Camilleri (liberato.camilleri@um.edu.mt)	Department of Statistics University of Malta, Malta
Line Harder Clemmensen (line.clemmensen@mail.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Denmark
Lisbeth Carstensen (lisbeth@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Maja Olsbjerg Larsen (m02mol@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Marit Ulvmoen (marit.ulvmoen@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Marta Lisa Diaz (m02mld@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Mathias Lindholm (lindholm@math.su.se)	Department of Mathematics Stockholm University, Sweden
Mats Rudemo (rudemo@math.chalmers.se)	Chalmers University of Technology Gothenburg, Sweden
Merete Jacobsen (mgj@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Mette Langaas (Mette.Langaas@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Michael Höhle (hoehle@stat.uni-muenchen.de)	Department of Statistics University of Munich, Germany
Michael Sørensen (michael@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Momodou Bah (mebah2004@yahoo.com)	Banjul, Gambia
Niels Keiding (N.Keiding@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Niels Richard Hansen (richard@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Nina Breinegaard (s03nb@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Ole F. Christensen (olefc@birc.au.dk)	Bioinformatics research center Aarhus University, Denmark
Ole Olsen (ole.olsen@gpract.ku.dk)	Research unit of general practice Copenhagen, Denmark
Olle Häggström (olleh@math.chalmers.se)	Mathematical Sciences Chalmers University of Technology, Göteborg, Sweden

Name (email)	Affiliation
Per Bruun Brockhoff (pbb@imm.dtu.dk)	Informatics and Mathematica Modelling Technical University of Denmark, Denmark
Peter Lewy (pl@dfu.min.dk)	DFU Copenhagen, Denmark
Pia Larsen (p.v.larsen@stat.sdu.dk)	University of Southern Denmark Odense, Denmark
Randi Grøn (m02rg@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Rasmus Nielsen (rasmus@binf.ku.dk)	Centre for Bioinformatics University of Copenhagen, Denmark
Rasmus Theis Lange (rathla@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Rasmus Waagepetersen (rw@math.aau.dk)	Department of Mathematical Sciences Aalborg University, Denmark
René Tabanera y Palacios (rtp@novonordisk.com)	Novo Nordisk A/S Bagsværd, Denmark
Robin Henderson (Robin.Henderson@newcastle.ac.uk)	Mathematics and Statistics Newcastle University, United Kingdom
Rolf Sundberg (rolfs@math.su.se)	Mathematical statistics Stockholm University, Sweden
Rune Viig Overgaard (rvo@imm.dtu.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Copenhagen, Denmark
Sanne Gørtz (sgoertz@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Seppo Pynnönen (sjp@uwasa.fi)	University of Vaasa Finland
Ståle Nygård (staaln@math.uio.no)	Department of Mathematics University of Oslo, Norway
Susann Stjernqvist (susann@maths.lth.se)	Centre for mathematical sciences Lund University, Sweden
Susanne Ditlevsen (S.Ditlevsen@pubhealth.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Sven Ove Samuelsen (osamuels@math.uio.no)	Department of Mathematics University of Oslo, Norway
Svend Kreiner (s.kreiner@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark
Søren G. Erbou (sge@imm.dtu.dk)	Informatics and Mathematical Modelling Technical University of Denmark, Denmark
Søren Grimstrup (m03sg@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Søren Højsgaard (sorenh@agrsci.dk)	Danish Institute of Agricultural Sciences Tjele, Denmark
Søren Johansen (sjo@math.ku.dk)	Department of Applied Mathematics and Statistics University of Copenhagen, Denmark
Tatjana Pavlenko (tatjana.pavlenko@miun.se)	TFM, Mid Sweden University Sundsvall, Sweden
Terry Speed (terry@stat.berkeley.edu)	Department of Statistics University of California, Berkeley, USA
Thomas Scheike (ts@biostat.ku.dk)	Department of Biostatistics University of Copenhagen, Denmark

Name (email)	Affiliation
Thordis Linda Thorarinsdottir (disa@imf.au.dk)	The T.N. Thiele Centre University of Aarhus, Denmark
Tobias Rydén (tobias@maths.lth.se)	Centre for Mathematical Sciences Lund University, Sweden
Tom Britton (tom.britton@math.su.se)	Stockholm University Sweden
Torben Martinussen (torbenm@dina.kvl.dk)	Department of Natural Sciences Royal Veterinary and Agricultural University Copenhagen, Denmark
Trond Sagerup (sagerup@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Tue Tjur (tuetjur@cbs.dk)	Copenhagen Business School Denmark
Turid Follestad (turid.follestad@math.ntnu.no)	Department of Mathematical Sciences Norwegian University of Science and Technology Trondheim, Norway
Vanessa Didelez (vanessa@stats.ucl.ac.uk)	Statistical Science University College London, United Kingdom
Vilijandas Bagdonavicius (algirdasbag@techas.lt)	Vilnius university Vilnius, Lithuania
Yudi Pawitan (yudi.pawitan@ki.se)	Karolinska Institutet Stockholm, Sweden
Yuri Goegebeur (yuri.goegebeur@stat.sdu.dk)	Department of Statistics University of Southern Denmark, Odense, Denmark
Ørnulf Borgan (borgan@math.uio.no)	Department of Mathematics University of Oslo, Norway