Statistical Methods for Complex and High Dimensional Models

Michael Sørensen, professor
Department of Mathematical Sciences
Universitetsparken 5, DK-2100 Copenhagen Ø
E-mail: michael@math.ku.dk
Tel.: +45 3532 0680 Fax: +45 3532 0704

Submitted to
The Faculty of Science
The Faculty of Health Sciences
The Faculty of Life Sciences

Four potential evaluators:

George Casella, professor, Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, U.S.A. Phone: +1 352 392 1941 ext 204. Fax: +1 352 392-5175. E-mail: casella@stat.ufl.edu.

Ursula Gather, professor, Department for Mathematical Statistics with Application in Industry, University of Dortmund. Address: Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany. Phone: +49 231 755 3110. Fax: +49 231 755 5305.
E-mail: gatherstatistik.uni-dortmund.de.

Peter Green, professor, FRS, Department of Mathematics, University of Bristol, Bristol, BS8 1TW, U.K. Phone: +44 117 928 7967; Fax: +44 117 928 7999. E-mail: P.J.Green@bristol.ac.uk.

Peter McCullagh, professor, Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, IL 60637, U.S.A. Phone: +1 773 702 8340. Fax: +1 773 702 9810. E-mail: pmcc@galton.uchicago.edu.

Project description

Aims and visions

Today's information age revolves around collection and analysis of data. Digital technology has enabled easy collection and exchange of data, and the massive explosion in the amount of data and the level of details that has become available is a major challenge in the quantitative sciences. This is exemplified by the development in molecular biology, where it has become possible to measure quantitatively the amount of RNA-transcript for thousands of genes simultaneously. To take full advantage of huge and very detailed data sets, statistical methodology for complex and high-dimensional models must be developed, and this research is gaining momentum internationally. The computing capacity now available has made the use of statistical methods based on complicated models practicable, and such methods are becoming a vital component of many areas of natural and social sciences, for instance bioinformatics, finance, geophysics, and life and health sciences. These developments pose both an opportunity and an interesting challenge to the statistical science. The project aims to play a leading role in the development of statistical methodology for complex and high-dimensional models in a number of selected areas where the University of Copenhagen has particular strength and research activity at a high international level, and where, moreover, advantage can be taken of recent developments in probability theory, e.g. in stochastic calculus, in the analysis of the new statistical methodology. This is a natural continuation of a long and internationally highly recognized Danish tradition for development and analysis of statistical models that has had a significant impact on the applications of such models in practice. One successful example of this Danish tradition is survival analysis.

The merger of the University of Copenhagen and the Royal Veterinary and Agricultural University has brought together three strong statistics groups, one at the Faculty of Science, one at the Faculty of Health Sciences and one at the Faculty of Life Sciences, which individually have a prominent international position.

The vision of the project is to combine the strengths of the three statistics groups at the University of Copenhagen to create a centre of statistical science that will belong to the globally leading centres by using the strong mathematical background of all three groups and the synergistic effect of bringing together statisticians from applied and theoretical environments. The centre will develop and investigate new statistical methodology that is needed to analyse the vast amounts of complex data available in many quantitative sciences and in industry.

Research plan

The main scientific goal of the project is to develop statistical methodology and theory for large and complex data sets. Particular attention will be given to data sets of the following types: data sets with a large number of simultaneous, parallel measurements, possibly many explanatory variables, but few replications (gene expression is an example); high-dimensional, but possibly short, time series and longitudinal data (panel data); and data with a high level of detail (e.g. high frequency financial time series and spectroscopic data). Simple models are typically not adequate as the dependency structures in the variables monitored are complex, and model-classes that can capture such structures typically have a large number of parameters, which makes statistical inference difficult. Moreover, the structure of the data often has to be used in a systematic approach to the problems under study to avoid ad hoc analyses using basic classical statistical methodology resulting in numerous fragmented

analyses with unreliable spurious findings, as has been the case with analyses of microarray data until more statistical approaches emerged. The scope of the project is not only to develop and analyse new statistical models, but also to develop new directions of fundamental statistical theory.

Popular current approaches to deal with the high-dimensional problems include dimension reduction techniques and various shrinkage or regularization approaches. In the extreme situation the approaches reduce to pure function approximation that can be useful for prediction purposes, but is hard to interpret. By keeping the focus on statistical models and model analysis, it is intended to retain the interpretability and thus the qualitative understanding in the quantitative analysis – also of high-dimensional data.

Useful statistical theory can only be developed in an interaction with applications. Therefore *the project will focus on the following themes where project participants have already internationally highly recognized research activity and deep involvement with the areas of application.*

- Statistical aspects of bioinformatics and gene regulation
- Survival analysis
- Dynamical stochastic models
- Image analysis
- Functional data
- Statistical computing

A further reason for the selection of these themes is that for all of them there is research activity at two or more of the three participating faculties, so that the synergistic effect of bringing together statisticians from the three environments can more easily be realized.

Statistics in Denmark is internationally renowned for combining theoretical strength with profound and dedicated applications of statistics in other disciplines like public health, biology and econometrics. This strong balance is maintained at our university by having not only a group of mathematical statistics with a good view towards applications (at the Faculty of Science), but also groups of theoretically highly qualified statisticians solidly planted and well integrated in the local application environments (at the Faculty of Health Sciences and the Faculty of Life Sciences). This latter integration is crucial not only for the continuing deep and serious involvement with the applications, but - in the present context even more so - for the essential and inexhaustible source of problems that require new development of statistical methods and theory. Similarly, the location of mathematical statistics in the Department of Mathematical Sciences at Faculty of Science is essential for maintaining the high theoretical standard in research and in teaching of future statisticians.

The project researchers are

- Principal investigator: **Michael Sørensen**, professor, Dept. Math. Sciences, Faculty of Science.
- Scientific coordinator at Faculty of Health Sciences: **Niels Keiding**, professor, Dept. of Biostatistics, Faculty of Health Sciences.

- Scientific coordinator for bioinformatics: **Anders Krogh**, professor, Bioinformatics Centre, Faculty of Science.
- Scientific coordinator at Faculty of Life Sciences: **Ib Skovgaard**, professor, Dept. Natural Sciences, Faculty of Life Sciences.
- **Claus Ekstrøm**, associate professor, Dept. Natural Sciences, Faculty of Life Sciences.
- **Niels Richard Hansen**, assistant professor, Dept. Math. Sciences, Faculty of Science.
- **Thomas H. Scheike**, professor, Dept. of Biostatistics, Faculty of Health Sciences.

Bioinformatics is a research area of high priority in the project. Therefore the project researchers include Professor Anders Krogh from the Bioinformatics Centre as scientific coordinator for bioinformatics, and a very promising young researcher in the area of statistical bioinformatics Assistant Professor Niels Richard Hansen. The inclusion of Anders Krogh will ensure close collaboration with the Bioinformatics Centre at the Faculty of Science. Also close contact to the Centre for Applied Bioinformatics at the Faculty of Life Sciences is ensured since Associate Professor Claus Ekstrøm is head of this centre.

In the following the research themes will be described in detail.

Statistical aspects of bioinformatics and gene regulation

Gene expression data is a significant example of great practical importance of high-dimensional measurements from a complex system with structured dependencies. The expression levels of different genes are dependent for instance via shared regulatory mechanisms, but as the typical dataset consist of thousands of parallel expression measurements and only few replications, inference concerning the dependency structure is a difficult and challenging problem. However, some structure can be expected in the dependencies reflecting the underlying molecular mechanisms such as the organization of transcription factors and other regulatory mechanisms. As a first approach, a recently suggested shrinkage method for estimation of covariance matrices, Ledoit & Wolf (2003,2004), developed originally for use in portfolio optimization in finance, was found equally useful to infer the dependency structure for gene expression measurements, Schäfer & Strimmer (2005), thus illustrating the cross-disciplinary nature of statistics. Another recent promising approach to estimation of the covariance matrix was proposed by Bickel & Levina (2006).

The organization of genomes (here we focus on Eukaryotes) and in particular the potentially regulatory motifs, present a similar problem of a high-dimensional, complex system of measurements. The measurements, arising from a computational or experimental source, are the positions of the various genes and/or motifs and potentially also additional information such as a quantitative score. Again structure can be expected in the organization of the genes and motifs with dependencies in the positioning of different regulatory motifs, say, reflecting co-regulatory properties. A general class of statistical point process models to capture co-occurrence phenomena of motifs was suggested in Gusto & Schbath (2005).

The organization of the genome and the resulting dependency structure in the gene expression measurements are also expected to be tightly linked. An important research goal is therefore to link the data and models on the organizational level with the models used for gene expression data.

It is intended to first investigate the following three research problems:

1. Development and analysis of multivariate marked point process models with the genomic organization as the main application.
2. Development and analysis of structured covariance models with particular emphasis on situations with a large number of simultaneous measurements and few replications, the main application being inference of dependencies in gene expression measurements.
3. Integration of genome organization models and gene expression models. In particular, how to encode organizational structure into a suitable dependence structure for the gene expression measurements.

A starting point for the development of the multivariate marked point process models will be the class of Hawkes models as suggested in Gusto & Schbath (2005). A multivariate version of this class of models will allow modeling of several mutually interacting point processes of motif occurrences. In an intensity formulation it is possible to capture both up- and down-regulation of the occurrence of one motif given one or several other motifs. The use of spline-based expansions of the intensities will be investigated, and the intensity functions will be estimated by penalized maximum likelihood. A problem of particular relevance is to judge whether one motif affects the occurrence of another, which can be formulated as a (non-parametric) model restriction with a constant conditional intensity. A further question is adaptation of models without the "time orientation" that Hawkes processes have. Such models would typically be formulated as Gibbs or Cox processes in the spirit of spatial point processes, see Møller & Waagepetersen (2004). For such processes the likelihood is not as easily accessible and other inferential procedures may be needed. A short term goal for this research is to get a working class of models and methods that can effectively capture the influence of the occurrence of one regulatory motif on the occurrence of other motifs in the genome. A long term goal is to develop a flexible class of suitable multivariate marked point process models, to develop the corresponding statistical theory, and to get the models implemented in the framework of R/Bioconductor to be easily accessible to other researchers. It is expected that graphical models will be a useful way to integrate genome organization models in models of gene expression data by representing expected conditional independence structures in terms of graphs.

Survival analysis

High-dimensional covariates

The emergence of gene-expression and SNP data, which has dramatically increased the amount of subject specific data, poses an interesting challenge to the statistical techniques traditionally used for analysis of survival data. New methodology will be developed, studied and implemented for extracting information from high-dimensional covariates like gene-expression and SNP data that can be used to predict the survival of specific patients and thus provide a very useful tool for inference about genetic effects. A main challenge is that, in contrast to the standard situation, the number of patients is small compared to the number of explanatory covariates for each patient.

Suitable dimension reduction method must be found. The Partial Least Squares (PLS) method, popular in chemometrics, constructs a low-dimensional explanatory vector that explains most of the response variation, by means of repeated orthogonalization and least squares analyses. Such a technique, however, does not tie well in with event time data because the response may be censored (event-free). Rather than using least squares methods, an approach

that deals neatly with the censoring problem is to model the dependence of the hazard function on explanatory variables. It will be investigated how dimension reduction methods, such as the PLS, can be accommodated to hazard function regression models. PLS methods have been suggested in the context of the proportional hazards model (e.g. Li & Gui, 2004, Park et al., 2002, Nygård et al., 2006). This project will focus on the more convenient class of additive hazards models (Martinussen and Scheike, 2006), where the additive structure allows the development of a much simpler theory and computationally much simpler methods.

Also other techniques for dimension reduction will be developed in the context of additive hazard regression. The focus will be on model fitting criteria such as the LASSO (Hastie et al., 2001) and SCAD with the oracle property (Fan & Li, 2002) that will be developed in the context of the additive hazards models. These techniques have an appealing sparsity property such that a small number of genes will be identified as being of importance. Recent theoretical work on L_1-based methods shows that these methods are consistent even when the number of covariates increases with the sample size (Buhlmann, 2006, Fan & Peng, 2004, Zhou, 2006).

Recently procedures have been proposed that implement Cox's regression models for haplotype association studies in an efficient manner (e.g. Zeng et al., 2006). Alternative procedures will be considered that are easier to implement, and additional methodology will be developed for estimating a time-varying effect of haplo-type based on flexible hazard models such as the Cox-Aalen model.

Random effect models for survival data

The random effects modelling approach may be considered as an approach for dealing with high-dimensional covariates, an idea investigated by Pawitan et al., (2004). It is intended to continue previous work in the project group, described below, on random effect models and to develop methodology and software for fitting more complicated random effects models for survival data that can describe subject specific treatment effects and are based on theoretically justified procedures. This will be done using likelihood-based techniques as well as estimating equation approaches.

Some progress has been made on random effects models for survival data in the context of Cox's regression model starting with work on the frailty model (Andersen et al., 1993), but there is little work on non-parametric estimation of treatment effects in this context and this seriously limits the use of these models in practical work. The theoretical problem is that the suggested likelihood approach (Nielsen et al., 1992; Parner, 1998) involves an infinite dimensional parameter in terms of the baseline of Cox's regression model. Recently techniques and software have been developed for so-called marginal models for survival data. Based on the additive hazards model, Martinussen and Scheike (2006) have shown how to estimate marginal treatments effects non-parametrically.

Dynamical stochastic models

It is often most natural to model the dynamics of a stochastic system in continuous time where models can be formulated in terms of stochastic differential equations and/or hazard functions and can be analysed by methods from stochastic calculus. A continuous time formulation often implies that the statistical parameters are more readily interpretable in applications. Statistical methods will be developed and studied for the following process types.

Point processes

Point processes form the basis of the hazard function formulation of survival analysis and are

a crucial tool in the proposed bioinformatics project, so the study of point processes is an integral part of these other projects.

Markov jump processes

Statistical methods will be investigated for Markov jump processes where the jump intensity depends on other unobserved Markov jump processes. Models and methods will be developed in order to analyse credit rating data, where the ratings of a large number of firms are observed for a number of years. The data are a large number of relatively short time series of ratings. The dependence between different firms will be modelled in terms of the unobserved components of the model. Observations are not necessarily made continuously, but possibly at discrete time points, so that the data may be incomplete in two ways. This research will build on previous results in Bladt & Sørensen (2005, 2006)

Stochastic differential equations

The class of models given in terms of stochastic differential equations is very large and includes processes with jumps and memory (i.e. non-Markovian processes). Stochastic partial differential equations are defined in space and time and can be used to analyse time-dependent spatial data.

Models given in terms of stochastic differential equations cannot be observed continuously because the sample paths are too irregular. Observations are at discrete time points or can, for instance, be observations of integrals of the process. Observations can be incomplete in other ways too: there may be measurement error, or all variables needed in a meaningful multivariate model may not be observable. Methods based on estimating functions (Kessler & Sørensen, 1999, Bibby & Sørensen, 2001) as well as simulation-based Markov chain Monte Carlo methods (Bladt & Sørensen, 2007) will be considered. Previous work aimed at handling e.g. stochastic volatility models and integrated diffusions (Sørensen, 2000, Ditlevsen & Sørensen, 2004) provide an important basis for applying estimating functions.

Methods will be developed with the following types of data in mind. Financial time series (e.g. stock prices, interest rates or exchange rates) can be modelled by stochastic volatility models which can be viewed as partially observed multivariate stochastic differential equation models. Models used to analyse physiological data are of a similar type. It is more generally intended to develop statistical methods for stochastic differential equation models in systems biology and in particular of gene-regulation. Another goal is models and statistical methods for the analysis of ice core data in order to study the paleoclimate. Here integrals of the processes are observed. This part of the project will be done in collaboration with colleagues at the Niels Bohr Institute and the Humboldt-University of Berlin.

Non-linear time series

A different approach to modelling time series is to use models defined in discrete time. Such models are usually referred to as time series models. For the analysis of many economic time series it has turned out that non-linear models are needed. Non-linear time series models will be developed with a view to the analysis of macro-economic and financial time series. The project group has previously been very successful in developing methods for studying cointegration, i.e. to find stationary relations between non-stationary time series. In particular a non-linear theory of cointegration will be developed. It is also intended to used cointegration methods to study climate data.

Image analysis

Image analysis and error processes

High-resolution digital images are the source of many datasets for example from microarray experiments and electrophoresis gels. Statistical models are needed to identify and extract the information from such images – both in term of segmentation (determine which parts of the images that contain the signal) and extraction (estimating the observed signal). Often, artifacts are introduced by the technology or by the image scanner, but these artifacts are not completely random and statistical models can use the information about the experimental design to correct these artifacts and to identify the relevant image areas and extract the information.

It is not uncommon to extract information from images by superimposing a simple template showing the relevant parts of the image. These templates, however, often fail if artifacts occur in the image. Non-parametric methods such as PCR can be used to improve the image segmentation by accounting for variations and imperfections in the image. Non-linear models with serial correlation can improve the data extraction in some situations, and it is planned to investigate how well these parametric models can be used on low-quality image, where parts of the image is missing or the contrast is low. In analyses of 1D and 2D spectra smooth dislocations of peaks call for an alignment of the image as an initial step. A semi-Bayesian statistical fitting of parametrically modelled templates penalizes vast deformations without preventing them and provides a method of warping without the need of landmarks or segmentation of the image. Overlaying such a parametric warping with a small-scale diffusion allows for automatic fine-tuning of the warp.

Image analyses and covariates

The analysis of neuro-imaging data is enormously challenging due to the complexity of such data. The data essentially consist of a three-dimensional grid of volume elements (voxels) giving intensity levels that describe brain activity or molecular composition at different locations and possibly with measurements at several time points. This outcome must be related to covariates such as disease status and medication. The state-of-the-art technique is to apply standard tests at each voxel separately, which must then somehow be combined to form a conclusion. The analysis is given an additional layer of complexity when the functional brain data is related to a high-dimensional covariate such as SNPs or gene-expression data.

It is intended to develop global models and methodology including structural equation models and stochastic geometry modelling. Non-parametric techniques such as functional PCA might be used for dimension reduction.

Functional data

Alignment of spectra

Recent advances in technology have made the study of chemical fingerprints left by cellular processes very inexpensive and massive amounts of data are available from the metabolomic and proteomic fields. The data are separated using mass spectrometry or gel electrophoresis and the resulting functional data (often in the form of multi-dimensional spectra) are used to identify and quantify metabolites or proteins. However, identification of specific differences between spectra is a challenge – not only because of small sample size and ghosting, but also because non-linear shifts in one or more dimensions warp the image. Statistical models may be used to align images, either by use of well-determined landmarks or by automatic use of the common patterns in the images. In both cases the methods have the advantage of not

requiring band or spot identification before the alignment.

A smoothed version of the spectra is often used to align the observed data by landmark matching or (semi-)parametric transformation of the time axis. It is not always the general profile that is the main points of interest, but also the variance of the profile at a specific time point. In such situations, smoothed data is not a viable option for the subsequent analyses. We will investigate parametric transformations of the time axis that will lessen the requirement for smoothed data and allow for time-dependent variance.

Human gait

Traditional models summarize the human movement into continuous time recordings for the key joints such as pelvis, hip, knee and angle. Based on recordings from additional force plates that measure the force of the feet versus the ground one may describe the forces in various joints and muscles.

The main aim of this project is to be able to characterize differences in gait and thereby be able to compare various treatments as well as to separate subjects. When comparing gait in different groups it is important to correct for the effect of explanatory variables such as age and gender. Therefore a further aim is to extend the research to include regression modelling or conditional analysis.

The state-of-the-art is described in the book on functional data by Ramsay & Silverman (1997). Many challenges remain. We aim to study non-parametric techniques and to extend previous work on parametric modelling by Olshen, Biden, Wyatt & Sutherland (1989) by giving a more detailed description of different sources of variation.

Statistical inference in mixed linear model with serial correlation

The class of linear mixed models with serial correlation provides an extremely flexible framework for analysis of longitudinal data (long series of measurements taken on the same subjects) from many fields, e.g. dietary studies in human nutrition. The choice of correlation structure may greatly influence the conclusions of the statistical analysis. However, in many specific situations it can be difficult to choose between different serial correlation structures. Diggle (1988) proposes an informal graphical method for identifying the serial correlation structure. Alternatively a decision may rely on the maximum likelihood concept, either through the use of likelihood ratio tests or information criteria. In some situations (e.g. when testing the hypothesis of no serial correlation structure) the distribution of the likelihood ratio test statistic is non-trivial, and methods to evaluate the necessity of a serial correlation structure in such situations are based on Gaussian processes. Some work has been done in this area (Ritz and Skovgaard, 2005) and is ready for implementation, but further theoretical work is called for.

Statistical computing

In recent years, statistical computing has developed from being mainly an issue of implementing mathematical methods to a multi-facetted independent field in its own right with peer-reviewed and indexed journals. It is a particularly important field since it is the glue that binds theoretical methods to practical applications. It also plays a major role in the teaching of statistics at many levels.

A substantial element of current research in statistical computing involves fitting statistical methodology into the general computing landscape. Moreover, theoretical properties of computing languages for statistical analysis, such as the S language and its R dialect, and aspects of their run-time implementation deserve further study. Computational methods per se are also still an important topic in statistics, and there are quite a few "hard" computational

problems around. Careful consideration of computational complexity and efficient algorithms is also often needed. This is particularly true for the large data sets that are the subject of this proposal.

It is proposed to work on the following specific problems:

1. Higher-order approximations for mixed models and their connection to sparse-matrix based algorithms.
2. Generic likelihood methods, in particular working towards mixed-effects modeling and higher-order analysis.
3. Interfacing to symbolic mathematics software, and integration of symbolic mathematics in statistical modeling.

Project organization

Research within the main themes will run in parallel. The principal investigator will monitor the progress and ensure that advantage is taken of all potentials for collaboration and synergy. He will meet regularly with the three scientific coordinators, who will assist him in this.

Communication is very important in a inter-faculty project like this. All electronic means of communication will be utilized, including a project web-site and an electronic newsletter with news about the scientific progress and e.g. new hires. This will serve to enhance the team spirit among the project participants. It is intended to organize a joint project seminar and an annual internal project workshop.

Project potential

The project builds on a long and internationally highly recognized Danish tradition for development of statistical methods and theory based on a strong mathematical background, and equally important, on profound and dedicated applications of statistics in other disciplines. This tradition has had a significant impact on the applications of statistics in practice. Individually the statistics groups at three faculties have a strong international position and includes world leading experts in statistical science. The combination of the, to a large extent complementary, strengths of these groups will have the potential to produce outstanding research of international top quality and to create a centre that will be among the globally leading centres of statistical research.

Project relevance

The proposed research will create statistical methodology for large and complex data sets that is needed in many quantitative sciences and in industry. The results are thus also highly relevant for courses given to students of statistics and other mathematical sciences as well as to students of other sciences and from industry.

Potential synergistic effects

Strong synergistic effects are expected from bringing together statisticians from applied and theoretical environments who share a solid mathematical background that facilitates the proposed research collaboration. The research themes are selected to optimize the synergistic effect. The following table indicates at which of the three faculties there is now research activity in each of the six main themes.

	Science	Health	Life
Bioinformatics	x	x	x
Survival analysis	x	x	x
Stochastic dynamical models	x	x	
Image analysis		x	x
Functional data		x	x
Statistical computing	x	x	x

Moreover, large data sets is the unifying common problem that goes across the themes, and *similar ideas and methods can be used to solve problems within different themes.* Thus there is very considerable scope for synergy.

Communication strategy

The results of the joint project will be disseminated through the usual channels: The results will be presented at international conferences and workshops, and will be published in leading international scientific journals. It is, moreover, the intention to organize international workshops and summer schools at the University of Copenhagen on research themes related to the project, and a project web-site will inform about the obtained scientific results.

The results will, finally, be incorporated into courses taught to students of statistics and other mathematical sciences as well as to students of other sciences at the three faculties involved, and hopefully also at other faculties and at courses for participants from industry.

Plans for external funding

The principal investigator will participate in an application for a Marie Curie research training network under FP7 coordinated from the University of Amsterdam. It is planned that most of the themes in the present proposal will be included in the FP7 proposal. The principal investigator, moreover, has plans for an application for a center on dynamic stochastic models to the Danish National Research Foundation or the Danish Natural Science Research Council. If the present proposal is successful, the scope of this application will be broadened to include most or all of the themes in the present proposal. Other members of the project group plan to apply to the Danish Natural Science Research Council for funding of aspects of some of the research themes.

At the Faculty of Health Sciences and the Faculty of Life Sciences, approximately seven statistics positions out of eighteen and three out of eight, respectively, are financed externally through temporary funds from national and international, public and private research foundations (e.g. Research Councils, the Lundbeck Foundation, the Danish Environmental Protection Agency, the European Union, and the National Cancer Institute (USA)). It is intended to maintain this funding level. The funding possibilities are continuously assessed together with research partners from applied fields.

References

Andersen, P.K., Borgan, Ø., Gill, R., & Keiding, N. (1993): *Statistical models for counting*, Springer-Verlag, New York.

Bibby, B.M. & Sørensen, M. (2001): Simplified estimating functions for diffusion models with a high-dimensional parameter. *Scand. J. Statist.* **28**, 99 – 112.

Bickel, P. & Levina, E. (2006): Regularized estimation of large covariance matrices (Working paper).

Bladt, M. & Sørensen, M. (2005): Statistical inference for discretely observed Markov jump processes. *J. Roy. Statist. Soc., ser. B*, **67**, 395 – 410.

Bladt, M. & Sørensen, M. (2006): Efficient estimation of transition rates between credit ratings from observations at discrete time points. Preprint No. 2006-2, Department of Applied Mathematics and Statistics, University of Copenhagen.

Bladt, M. & Sørensen, M. (2007): Simple simulation of diffusion bridges with application to likelihood inference for diffusions. Statistics preprint No. 2007-2, Department of Mathematical Sciences, University of Copenhagen.

Buhlman, P. (2006): Boosting for high-dimensional linear models. *Ann. Statist.*, **34**, 559 – 583.

Diggle, P. (1988): An approach to the analysis of repeated measures. *Biometrics* **44**, 959 – 971.

Ditlevsen, S. & Sørensen, M. (2004): Inference for observations of integrated diffusion processes. *Scand. J. Statist.*, **31**, 417 – 429.

Fan, J. & Li, R. (2002): Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, **30**, 74 – 99.

Fan, J. & Peng, H. (2004): Nonconcave penalized likelihood with diverging number of parameters. *Ann. Statist.*, **32**, 928 – 961.

Gusto, G. & Schbath, S. (2005): FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Genet. Mol. Biol.*, **4**.

Hastie, T., Tibshirani, R., & Friedman, J. (2001): *The Elements of Statistical Learning*. Springer-Verlag, New York.

Kessler, M. & Sørensen, M. (1999): Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299 – 314.

Ledoit, O. & Wolf, M. (2003): Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance* **10**, 603 – 621.

Ledoit, O. & Wolf, M. (2004): A well conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365 – 411.

Li, H. & Gui, J. (2004): Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* **20**, 208 – 215.

Martinussen, T. & Scheike, T. (2006): *Dynamic Regression Models for Survival Data*. Springer-Verlag, New York.

Møller, J. & Waagepetersen, R.P. (2004): *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, Boca Raton.

Nielsen, G.G., Gill, R.D., Andersen, P., & Sørensen, T.I.A. (1992): A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25 – 43.

Nygård, S., Borgan, O., Lingjærde, O., & Størvold, H.L. (2006): Partial least squares Cox regression on genomic data handling additional covariates. Preprint, Dept. of Math.,

University of Oslo, 5, 1 – 20.

Olshen, R. A., Biden, E. N., Wyatt, M. P., & Sutherland, D. (1989): Gait analysis and the bootstrap. *Ann. Statist.*, **17**, 1419 – 1440.

Park, P., Tian, L., & Kohane, I.S. (2002): Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120 – 127.

Parner, E. (1998): Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, **26**, 183 – 214.

Pawitan, Y., Bjohle, J., Wedren, S., Humphreys, K., Skoog, L., Huang, F., Amler, L., Sharw, P., Hall, P., & Bergh, J. (2004): Gene expression profiling for prognosis using Cox regression. *Statist. Med.*, **23**, 1767 – 1780.

Ramsay, J. & Silverman, B. W. (1997): *Functional data analysis.* Springer-Verlag, New York.

Ritz, C. & Skovgaard, I.M. (2005): Likelihood ratio tests in curved exponential families with nuisance parameters present only under the alternative. *Biometrika* **92**, 507 – 517.

Schäfer, J. & Strimmer, K. (2005): A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology* **4** (1).

Stablein, D.M. & Koutrouvelis, I. A. (1985): A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, **41**, 643 – 652.

Sørensen, M. (2000): Prediction-based estimating functions. *Econometrics Journal*, **3**, 123 – 147.

Sørlie, T., Tibshirani, R., Parker, J., Hatie, T., Marron, J., Nobe,A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Peour, C.,Lønning, P., Brown, P., Børresen-Dale, A., & Botstein, D. (2003): Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, **100**, 8418 – 8423.

Zeng, D., Lin, D.Y., Avery, C.L., North, K.E., & Bray, M.S. (2006): Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics*, **7**, 486 – 502.

Zhou, H. (2006): The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418 – 1429.