The FocuStat Conference

Vårens Vakreste Variabler

May 22–25 2018 **Ingeniørenes Hus, Oslo**

focustat
FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

UiO : University of Oslo

The Research Council of Norway

The structure of the workshop is intended to encourage and facilitate active discussion, during, after and between talks. The time schedule indicated here is therefore not necessarily followed strictly.

Tuesday

8:45 - 9:15 Good morning & tea and coffee

9:15 - 9:30 **Nils Lid Hjort**:
Welcoming remarks: FocuStat; general themes; looking ahead

Block I: Confidence Distributions and Related Themes

9:35 - 10:15 **Jan Hannig:**
Deep Fiducial Inference

tea & coffee

10:40 - 11:15 **Tore Schweder:**
Unbiased confidence

11:20 - 12:00 **Leiv Tore Salte Rønneberg:**
Confidence distributions and objective Bayesian inference – some comparisons

lunch

Block II: Survival and Event Histories

13:30 - 14:10 **Riccardo de Bin:**
Strategies to derive combined prediction models using both clinical predictors and high-throughput molecular data

tea & coffee

14:30 - 15:10 **Håkon Gjessing:**
Joint modeling of fetal size and time-to-birth

15:15 - 15:55 **Nils Lid Hjort:**
Survival and event history analysis via gamma process models

Wednesday

Block III: FICology

9:15 - 9:55 **Martin Jullum**:
Parametric or nonparametric, that's the question

tea & coffee

10:30 - 11:10 **Gerda Claeskens**:
Keep the focus

11:15 - 12:00 **Vinnie Ko**:
FIC for copulae

lunch

Block IV: High-dimensional models and applications

13:00 - 13:40 **Sylvia Richardson**:
Uncovering structure in high dimensional data via outcome-guided clustering

tea & coffee

14:00 - 14:40 **Aliaksandr Hubin**:
Deep Bayesian regression models

14:45 - 15:30 **Kristoffer Hellton:**
High-dimensional asymptotics of principal component regression

Thursday

Block V: From Processes to Models

9:15 - 9:55 **Alex Whitmore**:
Distribution-free inference methods for threshold regression

tea & coffee

10:30 - 11:10 **Emil Aas Stoltenberg**:
Models and inference for on-off data via clipped Ornstein-Uhlenbeck processes

11:15 - 12:00 **Per Mykland and Lan Zhang**:
The five trolls under the bridge: Principal Component Analysis with asynchronous and noisy high frequency data

lunch

Block VI: Dynamics, Change-points, Complexities

13:00 - 13:40 **Idris Eckley**:
Anomalies, change-points and exoplanets

tea & coffee

14:00 - 14:40 **Gudmund Hermansen**:
Bayesian nonparametrics for time series

14:45 - 15:30 **Ingrid Hobæk Haff**:
Focused selection of the claim severity distribution in non-life insurance

Friday

Block VII: Cool Strong Application Stories

9:00 - 9:50 **Ørnulf Borgan**:
Do Japanese and Italian women live longer than women in Scandinavia?

9:55 - 10:40 **Arnoldo Frigessi**:
Personalized computer simulation of breast cancer treatment: A multiscale dynamic model informed by multi-source patient data

tea & coffee

10:55 - 11:35 **Håvard Nygård**:
Why (and when) do small conflicts become big wars?

11:40 - 12:20 **Céline Cunen**:
Whales, politics, and statisticians

lunch

FocuStat, Focus Driven Statistical Inference With Complex Data, is a five-year project funded in part by the Research Council of Norway, operating from January 2014 to December 2018 at the Department of Mathematics, University of Oslo. The core of the project group consists of Nils Lid Hjort (professor, project leader), Gudmund Hermansen and Kristoffer Hellton (Post-Docs), Céline Cunen, Sam-Erik Walker, Vinnie Ko and Emil Aas Stoltenberg (PhDs). Other PhD and Master's level students are also associated with the project, and we are collaborating with yet other colleagues, at the Department of Mathematics and elsewhere.

The themes of the project include and involve confidence distributions; model selection and model averaging; bridging the gaps between parametric, semiparametric and nonparametric modelling and inference; Bayesian nonparametrics; combination of information across diverse data sources; survival and event history analysis; robustification of likelihood based inference methods; constructing models for data via background processes; etc.

A common thread is the notion of *focus*, the view that some aspects of experiments, data, and information are more important than others, and that such a science- and context-driven focus ought to contribute to the modelling and analysis of data, as well as to the performance evaluation of the relevant methods. This leads to focus driven model building and model selection, etc. The project is meant to develop relevant parts of general statistical methodology but has also involved actual applications to the analysis of real-world complex data. For further information, regarding publications, talks, 'who we are', news and events, etc., consult the project webpage `https://www.mn.uio.no/math/english/research/projects/focustat/index.html` and check (and contribute to) our Facebook page.

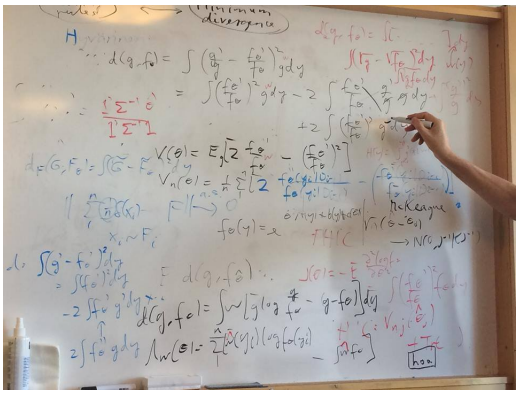
FocuStat has had budget for arranging moderate-scale international **workshops** on designated themes, for the spring semesters of 2015, 2016, 2017, as well as for lower-scale **research kitchens**, for the autumn semesters of 2014, 2015, 2016, 2017, 2018.

The present **Vårens Vakreste Variabler** conference of **May 2018** is a bit bigger in scope than the three-day theme based workshops and kitchens, and we are blessed by having about half of the University of Oslo's honorary doctorate winners in statistics, and about sixty percent of the Sverdrup Prize Winners, as speakers. Some of our earlier workshop themes are revisited, along with new types of applications. In addition to checking and using the titles *&* abstracts part of the present programme, see below for briefly keyworded summaries of our earlier workshops and kitchens.

The **May 2015** workshop **Inference With Confidence** concerned confidence distributions and related themes. These include and involve construction methods for CDs, studying their behaviour and performance, second-order correction tools for enhancing accuracy, links to so-called objective Bayes and empirical Bayes methods, meta-analysis and more general methods for data fusion, along with applications to real data stories. A Special Issue of *Journal of Statistical Planning and Inference* has been published in 2017, partly based on invited talks for the 2015 workshop, and with N.L. Hjort and T. Schweder as guest editors. The speakers were C. Cunen, A. Frigessi, S. Grønneberg, J. Hannig, K. Hellton, G. Hermansen, N.L. Hjort, B. Lindqvist, R. Liu, T. Schweder, D. Sun, G. Taraldsen, P. Veronese, S.-E. Walker, M.-ge Xie.

The **May 2016** workshop was on **FICology**, concerning focused ways in which to build and select models, along with model averaging, post-selection issues, and more. Methods and applications involve uses and variations of Focused Information Criteria. The speakers were A. Charkhi, G. Claeskens, C. Cunen, A. Gandy, P. Grünwald, S. Grønneberg, I. Hobæk Haff, H. Hegre, G. Hermansen, K. Hellton, N.L. Hjort, M. Jullum, I. Van Keilegom, T. van Ommen, E. Pircalabelu, S.-E. Walker, L. Walløe, P. Østbye.

The **May 2017** workshop was on **Building Bridges**, broadly speaking having to do with inventing, assessing, examining and fruitfully utilising ways of interconnecting parametrics, semiparametrics and nonparametrics. Themes include constructing nonparametric envelopes around parametric models (via Bayesian nonparametrics or otherwise), models with growing complexity, nonparametrics corrections to parametric pilot estimators, averaging across different types of models, advanced model selection, data fusion with parametric and nonparametric components, etc. Here the speakers were C. Cunen, R. de Bin, T. Egeland, I. Glad, K. Hellton, G. Hermansen, N.L. Hjort, M. Jullum, V. Ko, J. Moss, H. Otneim, S. Petrone, I. Prünster, C. Rohrbeck, B. Støve, E. Stoltenberg, D. Tjøstheim, S.-E. Walker.



Real-time kitchening.

The FocuStat group.

We were barely started in **August 2014**, but we had time for a small kitchen on **empirical likelihood** with a few delightful digressions; the 'hybrid likelihood' (Hjort, McKeague, Van Keilegom) came out of this kitchen.

In **September 2015** our kitchen was on **Minimum divergence methods** (including also proper scoring rules, prediction methodology, and of course bits of minimum dispair, maximum despair). Speakers were C. Cunen, K. Hellton, N.L. Hjort, F. Krüger, M. Musio, T. Thorarinsdottir, S.-E. Walker.

In **October 2016** our kitchen theme was **L$^\eta$ многая лета**, with the many and partly related ways in which one can work with robustified likelihoods, including constructions involving an extra exponent. Speakers were T. Broderick, I. Glad, P. Grünwald, K. Hellton, G. Hermansen, N.L. Hjort, J. Miller, P. Müller, P. Mykland, S.-E. Walker,

The **November 2017** kitchen themes were those of **From Processes to Models**. How can the processes behind the observed data be modelled, and with which consequences, even in cases where the background mechanism cannot be observed themselves? Speakers were R. de Bin, S. Engebretsen, C. Cunen, C. Heinrich, G. Hermansen, K. Hellton, N.L. Hjort, B. Lindqvist, J. Moss, R. Parviero, E. Stoltenberg, T. Thorarinsdottir, M. Tveten, S.-E. Walker.

Finally, for **November 2018** the kitchen plan appropriately involves mixing of ingredients and new ways of combining such for new models and new insights – how to fuse statistical information across diverse sources. Contact Céline and Nils if you might be interested in taking part.

Titles & abstracts:

Riccardo de Bin:

Strategies to derive combined prediction models using both clinical predictors and high-throughput molecular data

In biomedical literature, numerous prediction models for clinical outcomes have been developed based either on clinical data or, more recently, on high-throughput molecular data (omics data). Prediction models based on both types of data, however, are less common, although some recent studies suggest that a suitable combination of clinical and molecular information may lead to models with better predictive abilities. This is probably due to the fact that it is not straightforward to combine data with different characteristics and dimensions (poorly characterized high-dimensional omics data, well-investigated low-dimensional clinical data). Here we show some possible ways to combine clinical and omics data into a prediction model of time-to-event outcome. Different strategies and statistical methods are exploited.

Ørnulf Borgan:

Do Japanese and Italian women live longer than women in Scandinavia?

Life expectancies at birth are routinely computed from period life tables. When mortality is falling, such period life expectancies will typically underestimate real life expectancies, that is, life expectancies for birth cohorts. Hence, it becomes problematic to compare period life expectancies between countries when they have different historical mortality developments. For instance, life expectancies for countries in which the longevity improved early (like Norway and Sweden) are difficult to compare with those in countries where it improved later (like Italy and Japan). To get a fair comparison between the countries, one should consider cohort data. Since cohort life expectancies can only be computed for cohorts that were born more than a hundred years ago, we suggest that for younger cohorts one may consider the expected number of years lost up to a given age. Contrary to the results based on period data, our cohort results then indicate that Italian women may expect to lose more years than women in Norway and Sweden, while there are no indications that Japanese women will lose fewer years than women in Scandinavia. The large differences seen for period data may just be an artefact due to the distortion that period life tables imply in times of changing mortality.

(Joint work with Nico Keilmahn.)

Gerda Claeskens:

Keep the focus

It is 15 years after the publication of the focused information criterion. While at first the idea of setting a focus and selecting the model that estimates this focus well was thought-provoking, it now is being used in several contexts. The most commonly used definition of 'estimating well' is that the estimator has a small mean squared error. In this talk I will explain why mean squared error is a useful concept in model selection, and how one can estimate the mean squared error using local misspecification. Despite the often repeated advertisement of lasso-type methods being able to simultaneously select variables and estimate parameters in high-dimensional regression models, also for such models, it is advantageous to think first about which focus to estimate, and then perform the selection by taking that estimator of the focus with a small estimated mean squared error. The high-dimensional data setting will be discussed and illustrated with some applications. A brief summary will be given about what we have learned about focused selection, what is currently ongoing and what are possible venues for the future.

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion [with discussion]. *JASA*.

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press.

Céline Cunen:

Whales, politics, and statisticians

Statistical questions sometimes find themselves at the centre of heated, politicised debates. A seemingly innocuous hypothesis about the body condition of Antarctic minke whales has been discussed for about ten years in the Scientific Committee of the International Whaling Commission, with most of the discussion concerning statistical issues. The main research question revolves around the claim that the body condition of the whales, as measured by for example fat weight, has decreased during the study years. We will describe the biological background and our analysis, including a new focused information criterion for linear mixed effect models. We will also reflect on the responsibility of the statistician when it comes to communications of assumptions and uncertainty, and share some thoughts on the complex and slightly absurd world of international political organisations.

(Joint work with Nils Lid Hjort and Lars Walløe.)

Cunen, C. and Hjort, N.L. (2017). Whales, politics, and statisticians. FocuStat Blog Post.

Cunen, C., Walløe, L. and Hjort, N.L. (2018). Focused model selection for linear mixed models, with an application to whale ecology. Submitted for publication.

Idris Eckley:

Anomalies, Change-points and Exoplanets

Anomaly detection is of ever-increasing importance for many applications, primarily due to the abundance of sensors within contemporary systems and devices. Such sensors are capable of generating a large amount of data, necessitating computationally efficient methods for their analysis. To date, much of the statistical literature has been concerned with the detection of point anomalies, whilst the problem of detecting anomalous segments – often called collective anomalies – has been relatively neglected. We will introduce work that seeks to address this gap by introducing a linear time algorithm based on a parametric epidemic change point model. We present an approach that, with provable guarantees, is able to differentiate between both point anomalies and anomalous segments. Our computationally efficient approach is shown to be better than current methods, and we demonstrate its usefulness on the challenging problem of detecting exoplanets using data from the Kepler telescope.

(Joint work with Alex Fisch and Paul Fearnhead.)

VVV back in the days. *(Oslo Museum)*

Arnoldo Frigessi:

Personalized computer simulation of breast cancer treatment: A multiscale dynamic model informed by multi-source patient data

Mathematical modelling and simulation have emerged as a potentially powerful, time- and cost effective approach to personalised cancer treatment. In order to predict the effect of a therapeutic regimen for an individual patient, it is necessary to initialize and to parametrize the model so to mirror exactly this patient's tumor. In this talk I present a new comprehensive approach to model and simulate a breast tumor treated by two different chemotherapies in combination or not. In the multiscale model we represent individual tumor and normal cells, with their cell cycle and others intracellular processes (depending on key molecular characteristics), the formation of blood vessels and their disruption, extracellular processes, as the diffusion of oxygen, drugs and important molecules (including VEGF which modulates vascular dynamics). The model is informed by data estimated from routinely acquired measurements of the patient's tumor, including histopathology, imaging, and molecular profiling. We implemented a computer system which simulates a cross-section of the tumor under a 12 weeks therapy regimen. We show how the model is able to reproduce four patients from a clinical trial, both responders and not. As an example we show by scenario simulation, that other drug regimens might have led to a different outcome.

This is one of the first multiscale mathematical models for a solid tumor, which incorporates discrete and continuous pharmacokinetics and pharmacodynamics factors, is mechanistic in nature, and informed by individual patient data.

(This is joint work with: Xiaoran Lai, Alvaro Khn Luque, Vessela Kristensen, Olav Engebråten, Therese Seierstad, Thomas Fleisher, Oliver Geier, Øysten Garred, Elin Borgen, Marie E. Rognes, Simon W. Funke.)

Håkon Gjessing:

Joint modeling of fetal size and time-to-birth

From around week 24 of pregnancy, the weight of a human fetus is frequently estimated using ultrasound. However, very few children get born that early, and those who do are pathological and may not represent the unborn children very well. Relatedly, a fetus that has a growth restriction may be small and also end up being born very early. How can a weight estimation model be properly calibrated to represent children in normal pregnancies? Ultrasound measurements can be seen as irregular measurements of an underlying growth process, and birth can be seen as a time-to-event outcome. I demonstrate how to combine the information to obtain good weight estimates under reasonable assumptions.

Jan Hannig:

Deep Fiducial Inference

R. A. Fisher, the father of modern statistics, developed the idea of fiducial inference during the first half of the 20th century. While his proposal led to interesting methods for quantifying uncertainty, other prominent statisticians of the time did not accept Fisher's approach as it became apparent that some of Fisher's bold claims about the properties of fiducial distribution did not hold up for multi-parameter problems. Beginning around the year 2000, the authors and collaborators started to re-investigate the idea of fiducial inference and discovered that Fisher's approach, when properly generalized, would open doors to solve many important and difficult inference problems. They termed their generalization of Fisher's idea as generalized fiducial inference (GFI). The main idea of GFI is to carefully transfer randomness from the data to the parameter space using an inverse of a data generating equation without the use of Bayes theorem. The resulting generalized fiducial distribution (GFD) can then be used for inference. After more than a decade of investigations, the authors and collaborators have developed a unifying theory for GFI, and provided GFI solutions to many challenging practical

problems in different fields of science and industry. Overall, they have demonstrated that GFI is a valid, useful, and promising approach for conducting statistical inference.

In this talk we discuss how certain computations within generalized fiducial framework can be made using deep network. The resulting approximation to the fiducial distribution is termed deep fiducial distribution (DFD). We conclude by summarizing several difficult open problems related to this approach.

(Joint work with Gang Li.)

Fisher, R.A. (1930). Inverse Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 528–535.

Hannig, J., Iyer, H., Lai, C.S. and Lee, T.C. (2016). Generalized Fiducial Inference: A Review and New Results. *JASA*, 1341–1361.

Ingrid Hobæk Haff:

Focused selection of the claim severity distribution in non-life insurance

Estimating the reserve is one of the main applications of the total loss distribution of a non-life insurance portfolio. The choice of claim severity distribution should therefore reflect this. Therefore, we have explored how the focused information criterion, FIC, aimed at finding the best model for estimating a given parameter of interest, the focus parameter, works as a tool for selecting the claim size distribution. As the reserve cannot be used directly as a focus parameter, we have tried different proxy focus parameters. To see how the FIC performs in this setting, compared to the other commonly used model selection methods AIC and BIC, we have performed a simulation study. In particular, we wanted to investigate the effect of the heavyness of the tail of the claim size distribution and the amount of available data. The performance of the different model selection methods was then evaluated based on the quality of the resulting estimates of the reserve. Our study shows the best of the focused criteria is the FIC_ϵ, based on one single quantile from the claim severity distribution. Further, the performance of the FIC_ϵ is mostly either comparable to or considerably better than that of the BIC, which is the best performing of the state of the art approaches. In particular, the FIC_ϵ works well when the data are heavy-tailed, when the sample size is rather low and when the parameter of interest is a quantile far out in the tail of the total loss distribution

Kristoffer Hellton:

High-dimensional asymptotics for principal component regression

Principal component analysis (PCA) is used to reduce the data dimension in a range of applications, from genetics to climatology and finance. The method constructs a low-dimensional set of highly informative surrogate variables, called principal components (PCs). The scores of the PCs represent the original data and are used as inputs in linear regression. However, PCA has been shown to be inconsistent in high dimension, in particular when the number of variables exceeds the sample size. Thus regression based on principal components will be inconsistent in high dimension as well. But if the eigenstructure is assumed to be dense, Hellton and Thoresen (2017) have shown that, as the dimension increases, the sample PC scores converge to a scaled and rotated version of the population scores. The asymptotic estimation error will, therefore, consist of a random scaling and rotation. I will present some ideas of how to utilize this structure to improve regression based on principal component scores.

Hellton, K.H. and Thoresen, M. (2017). When and why are principal component scores a good tool for visualizing high-dimensional data? *Scandinavian Journal of Statistics* 44, 581–597.

Gudmund Hermansen:

Bayesian nonparametrics for time series

There exist various parametric and nonparametric modelling strategies for stationary time series. For most parametric approaches it is fairly easy to apply Bayesian techniques, where the large-sample behaviour is well understood, associated with Bernshteĭn–von Mises theorems. It is not clear how to proceed with a Bayesian nonparametric approach, however. Also, establishing good large-sample behaviour for nonparametric constructions, like posterior consistency or Bernshteĭn–von Mises theorems, becomes much more challenging. Here we will work with a certain class of sample size dependent priors that model the spectral density of the time series as constant over an increasingly refined partition of the frequency domain. This enables Bayesian nonparametric analysis of stationary Gaussian time series. Furthermore, inference is easily carried out via approximations or simulations, and we are also able to establish precise and sufficient conditions needed to obtain Bernshteĭn–von Mises type of results for this particular construction.

(Joint work with Nils Lid Hjort.)

Hermansen, G.H. and Hjort, N.L. (2015). Bernshteĭn–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. *Journal of Statistical Planning and Inference* 166, 138–157.

Nils Lid Hjort:

Survival and event history analysis via Gamma process models

Many models applied to survival and life history analysis data are 'off-the-shelf', without paying much attention to the underlying processes involved. Even though several of these much-used models often work well, from exponential and Weibull regression to the proportional hazards of Cox, there is scope for improvement, and for learning more about the processes involved, by attempting to model more of the data generating processes themselves. In my talk I will use Gamma processes for such purposes. Model building construction A is to consider the time such a process needs to reach a threshold a as a life-time. This leads to a large but manageable class of distributions, where covariates may influence the thresholds in question, or the motors driving the Gamma processes, or both. A different model construction B takes an individual to be alive as long as the shocks or jumps of the process are below a threshold b. This leads again to classes of biologically plausible models for survival and event history data with covariates. I will also briefly point to extensions to competing risks setups, to recurrent events, to multi-stage illness processes, and to cure models. Applications to real data will be discussed.

(Joint work with Céline Cunen.)

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press.

Aliaksandr Hubin:

Deep Bayesian regression models

Regression models are addressed for inference and prediction in a wide range of applications providing a powerful scientific tool for the researchers and analysts coming from different fields. In most of these fields more and more sources of data are becoming available introducing a variety of hypothetical explanatory variables for these models to be considered. Model averaging induced by different combinations of these variables becomes extremely important for both good inference and prediction. Not less important, however, seems to be the quality of the set of explanatory variables to select from. It is often the case that linear relations between the explanatory variables and the response are not sufficient for the high quality inference or predictions. Introducing non-linearities and complex functional interactions based on the original explanatory variables can often significantly improve both predictive and inferential performance of the models. The non-linearities can be handled by deep learning models.

These models, however, are often very difficult to specify and tune. Additionally they can often experience over-fitting issues. Random effects are also not incorporated in the existing deep learning approaches. In this paper we introduce a class of deep Bayesian regression models with latent Gaussian variables generalizing the classes of GLM, GLMM, ANN, CART, logic regressions and fractional polynomials into a powerful and flexible Bayesian framework. We then suggest algorithmic approaches for fitting them. In the experimental section we test some computational properties of the algorithm and show how deep Bayesian regression models can be used for inference and predictions in various applications.

Martin Jullum:

Parametric or nonparametric, that's the question

Should one rely on a parametric or nonparametric model when analysing a certain data set? This question cannot be answered by classical model selection criteria like AIC and BIC, due to the lack of a proper likelihood for the nonparametric model. In this talk, we present a focused information criterion (FIC) for selecting among a set of parametric models and a nonparametric alternative. It relies in part on the notion of a focus parameter, a population quantity of particular interest in each specific statistical analysis. The FIC compares and ranks candidate models based on estimated precision of the different model-based estimators for the focus parameters. An extension to an averaged focused information criterion (AFIC), taking multiple focus parameters into account, will also be outlined. For presentational simplicity we shall mainly work with the i.i.d. setting, but we also discuss extensions to other settings. In particular, we shall see that a special case of the AFIC applied to categorical data, casts new light to the classical Pearson chi-squared test.

(Joint work with Nils Lid Hjort.)

Jullum, M. and Hjort, N.L. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica*.

Jullum, M. and Hjort, N.L. (2018). What price semiparametric Cox regression?

Vinnie Ko:

FIC for copulae

Focused information criterion (FIC) (Claeskens and Hjort, JASA 2003, Jullum and Hjort, Sinica 2017) is a model selection criterion that can compare non-nested parametric models with a non-parametric alternative. FIC evaluates candidate models based on the estimated precision of the focus parameter, where focus parameter is a population quantity of interest in statistical analysis (e.g. upper-tail probability and copula parameter).

By extending the FIC methodology of Jullum and Hjort, we develop the focused information criterion (FIC) for copula models. The FIC for copula models covers the three most popular estimation schemes: maximum likelihood, two-stage maximum likelihood (also known as inference functions for margins), and pseudo maximum likelihood.

By using the FIC for copula, one can compare the performance of models that are estimated with different estimation schemes. This is one of the biggest advantage of the FIC and is not possible with traditional model selection criteria like AIC and BIC, since they evaluate the model in terms of Kullback–Leibler divergence from the data generating model. We demonstrate the behaviour of FIC by using a simulation study and via an application to real data.

(Joint work with Nils Lid Hjort.)

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion [with discussion]. JASA 98, 900–916.

Jullum, M. and Hjort, N. L. (2017). Parametric or nonparametric: the FIC approach. *Statistica Sinica*.

Per Mykland and Lan Zhang:

The five trolls under the bridge: Principal Component Analysis with asynchronous and noisy high frequency data

We develop a principal component analysis (PCA) for high frequency data. As in Northern fairly tales, there are trolls waiting for the explorer. The first three trolls are market microstructure noise, asynchronous sampling times, and edge effects in estimators. To get around these, a robust estimator of the spot covariance matrix is developed based on the Smoothed TSRV (Mykland, Zhang, Chen, 2017). The fourth troll is how to pass from estimated time-varying covariance matrix to PCA. Under finite dimensionality, we develop this methodology through the estimation of realized spectral functions. Rates of convergence and central limit theory, as well as an estimator of standard error, are established. The fifth troll is high dimension on top of high frequency, where we also develop PCA. With the help of a new identity concerning the spot principal orthogonal complement, the high-dimensional rates of convergence have been studied after eliminating several strong assumptions in classical PCA. As an application, we show that our first principal component (PC) closely matches but potentially outperforms the S&P 100 market index, while three of the next four PCs are cointegrated with two of the Fama-French non-market factors. From a statistical standpoint, the close match between the first PC and the market index also corroborates this PCA procedure and the underlying S-TSRV matrix, in the sense of Karl Popper.

(Joint work with D. Chen.)

Mykland, P.A., Zhang, L. and Chen, D. (2017). The algebra of two scales estimation, and the S-TSRV: high frequency estimation that is robust to sampling times. Conditionally accepted by Journal of Econometrics.

Håvard Nygård:

Why (and when) do small conflicts become big wars?

We examine the effect of UN Peacekeeping Operations (PKOs) on the intensity of violence seen in civil wars. Reducing the intensity of violence is a core objective of PKOs but we still lack a proper understanding of how PKOs affect the underlying latent conflict-process. Using event-level data for all civil wars from 1989 to the present, we develop a Bayesian hierarchical modelling framework that allows us to study how deploying a PKO affects the intensity, i.e. the escalatory (and de-escalatory) patterns of violence, in these conflicts.

Cunen, C., Hjort, N.L. and Nygård, H. (2018). Statistical sightings of better angels.

Sylvia Richardson:

Uncovering structure in high dimensional data via outcome-guided clustering

Although the challenges presented by high dimensional data in the context of regression and subset selection are well-known and the subject of much current research, comparatively less interest has been paid to this issue in the context of clustering. In the setting that we consider, we have a large number of covariates as well as an outcome of interest, and the key challenge is to identify a subset of the covariates that provides a stratification of the population that is 'relevant' with respect to the outcome. By relevant, we mean that the clustering and partitioning sought after is predictive of the outcome. For example, a task, which is of great interest in precision medicine, is to use multi-omics data to discover subgroups of patients with distinct molecular phenotypes and clinical outcomes, thus providing the potential to target treatments more accurately.

Identifying relevant subgroups and the covariates characterising these can be particularly challenging when dealing with high-dimensional datasets, as there may be many covariates that provide no information whatsoever about population structure, or – perhaps more challenging – in which there may be covariate subsets that define clear stratification that is not useful in prediction the outcome. For example, when dealing with genetic data, there may be some

genetic variants that allow us to group patients in terms of disease risk, but others that would provide completely irrelevant stratifications (e.g. which would group patients together on the basis of eye or hair colour).

Bayesian profile regression is a semi-supervised model-based clustering approach that makes use of a response in order to guide the clustering toward relevant stratifications (Molitor et al. 2010, Papathomas et al. 2012). Here we consider how this approach can be extended to a 'multiview' setting, in which different groups of covariates ('views') define different stratifications, thus including an informative variable selection step, which allows in particular to separate the relevant view and predictive clustering structure from other structures present in the data. We will outline our Bayesian approach based on Dirichlet Processes and discuss its computational challenges. We will show some results in the context of breast cancer subtyping, which illustrate how the approach can be used to perform integrative clustering of multiple 'omics datasets.

(Joint work with Dr Paul Kirk, MRC Biostatistics Unit, University of Cambridge.)

J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010). Bayesian profile regression with an application to the National Survey of children's Health, *Biostatistics* 11, 484–498.

M. Papathomas, J. Molitor, C. Hoggart, D. Hastie and S. Richardson (2012). Exploring Data From Genetic Association Studies Using Bayesian Variable Selection and the Dirichlet Process: Application to Searching for Gene×Gene Patterns. *Genetic Epidemiology* 36, 663–674.

Leiv Tore Salte Rønneberg:

Fiducial and Objective Bayesian inference – some comparisons

In Bayesian analysis, the prior distribution is meant to capture the current state of knowledge about the parameters in a certain model. Once new data is gathered, this distribution is updated to a posterior distribution by way of Bayes' theorem. In absence of good prior information, the tradition had been to utilise uniform priors to represent a state of ignorance about the parameter values – a practice deemed "fundamentally false and devoid of foundation" by R.A. Fisher in 1930. Instead, he proposed the fiducial distribution, which he felt represented an objective posterior distribution of the parameters without the introduction of unwarranted prior information.

While the fiducial distribution fell into disfavour after Fisher's death, there emerged a school of objective Bayesian analysis, where the prior distribution is derived on the basis of formal rules instead of actual knowledge and experience. And in the past decade or so, the theory of fiducial inference has been revived as well. The goal of this talk is to discuss some of the ideas behind the objective Bayesian and fiducial approaches, and make some comparisons of the resulting distributions across some old and new examples.

Tore Schweder:

Unbiased confidence

A confidence curve $\mathrm{cc}(\theta)$ is unbiased whenever $\mathrm{cc}(\theta_0) \leq_{\mathrm{st}} \mathrm{cc}(\theta_1)$ for a false value of the parameter, θ_1, and θ_0 being true; here $\leq_{\mathrm{st}}$ means 'stochastically smaller than'. This concept generalizes unbiasedness in hypothesis testing and confidence set estimation. The archetypal confidence distribution $C(\theta, T) = 1 - F(T, \theta)$, understood as confidence curve for left open intervals, is trivially unbiased. Two-sided tail-symmetric confidence curves need not be unbiased. Confidence curves for the mean parameter μ in the normal model are unbiased. Confidence curves based on the deviance are unbiased asymptotically in smooth models. I conjecture that deviance-based confidence curves are unbiased in general in parametric models.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability*. Cambridge University Press.

Emil Stoltenberg:

Models and inference for on-off data via clipped Ornstein–Uhlenbeck processes

Most people are in good health most of the time and, unfortunately, sick some of the time. In this paper we study a model where for a sample of individuals whose health statuses are governed by independent latent Ornstein–Uhlenbeck process. A person is sick if the process is above a certain threshold, and in good health otherwise, and it is only this only 'clipped' zero-one version of the process that is actually observed. Moreover, the sample is not continuously monitored, and the health status of an individual is only ascertained at certain points in time. These time points might be different for each of the individuals, they need not be equidistant, or they might be generated according to a stochastic process (independent of the underlying OU-processes).

A clipped Gaussian process is no longer Markov (Slud 1989) and likelihood inference is not straightforward. We propose and study a version of the quasi-likelihood (Hjort and Varin 2008) for inference on the parameters governing the underlying OU-processes. Large-sample properties of the estimators are presented, and the methods are applied to a data set of 926 Brazilian children followed up over a period of 15 months. The results of this analysis are compared and contrasted with the counting process approach of Borgan et al. (2007) to the same data set.

(Joint work with Nils Lid Hjort.)

Borgan, Ø., Fiaccone, R.L., Henderson, R., Barreto M.L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. Scandinavian Journal of Statistics 34, 53–69.

Hjort, N.L., Varin C. (2008). ML, PL, QL in Markov Chain Models. Scandinavian Journal of Statistics 35, 64–82.

Slud, E. (1989). Clipped Gaussian processes are never M-step Markov. Journal of Multivariate Analysis 29, 1–14.

Alex Whitmore:

Distribution-free inference methods for threshold regression

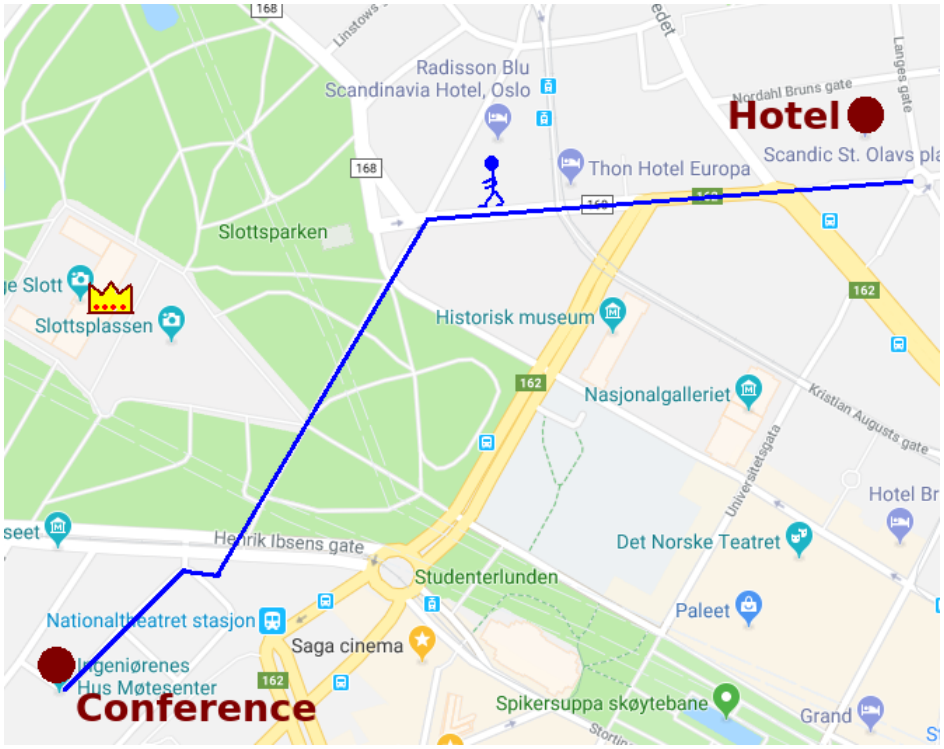
In many medical and health-care contexts, a failure event (such as death, hospitalization or transplant) is triggered when a subjects deteriorating health first reaches a failure threshold. The failure process is well described as the sample path of a stochastic process hitting a boundary. The parameters and behaviors of such failure processes must often be inferred from data sets that include censored survival times and current health levels of survivors. A substantial input of expert experience with the health context is usually required to guide the data modeling. This talk describes a parsimonious model for the failure process that has only one distributional property, namely, stationary independent increments. As this property is frequently encountered in real applications, the stochastic model and its related statistical methodology have potential for general application in many fields. The mathematical underpinnings of the distribution-free methods for estimation and prediction will be described in the talk as well as techniques for incorporating covariates. The methodology is essentially a distribution-free form of threshold regression. Computational aspects of the approach are straightforward. Case examples will be presented to demonstrate the methodology and its practical use. Several outstanding research questions of practical importance will be presented. The methodology provides medical researchers and analysts with new and robust statistical tools for assessing failure risks, estimating effects of risk factors and treatments, and making inferences about residual lifetimes of survivors. The methodology can help to deepen scientific insights into the causes and nature of disease progression.

(Joint work with Mei-Ling Ting Lee.)

Practical information:

Most of our international conference speakers will stay at Scandic St. Olavs plass, a 900 meters walk from Ingeniørenes Hus (the House of Engineers) where the Conference will take place. The address is Kronprinsens gate 17 and the nicest route is to walk across the Palace Park as indicated on the map below.

Both Ingeniørenes Hus and the hotel are close to the city centre and to several places of interest (if one has time for some sightseeing), with easy access by tram or the apostles' horses. The Royal Palace and Palace Park, as well as the Honorary cemetery 'Vår Frelsers gravlund', are peaceful places worth visiting. In the latter place one can find the graves of famous Norwegians like Henrik Ibsen, Edvard Munch, Bjørnstjerne Bjørnson, Henrik Wergeland, etc. If one wishes to admire some art by Munch, the National Gallery is found only 300 meters from the hotel. Some 200 meters further on, one finds the University Aula where Munch has decorated the walls with eleven fantastic paintings representing the different Sciences and the ideals of the Enlightenment.



Map of area between Scandic St. Olavs and Ingeniørenes Hus. The blue lines indicate the nicest walking route. *Google maps*

Conference dinner: On Wednesday the 23rd, the FocuStat group invites all speakers to join us for a social, cultural and gastronomical evening. We will walk together from St. Olavs plass at approximately 17:35. At 18:00 we will have a guided tour of Oslo City Hall. Most famous as the stage for the Nobel Peace Prize ceremony, it is also a great example of modernist architecture and filled with Norwegian art. Then, at 19:30, we will dine at the nearby restaurant Brasserie France, a popular French restaurant.

Participants:

The list is preliminary, and a few more are expected to take part, from the Department of Mathematics, the Norwegian Computing Centre and elsewhere.

From the FocuStat group: Céline Cunen, Gudmund Hermansen, Kristoffer Hellton, Nils Lid Hjort, Vinnie Ko, Emil Aas Stoltenberg, Sam-Erik Walker.

From the University of Oslo, Department of Mathematics: Josephina Argyrou, Riccardo de Bin, Ørnulf Borgan, Simon Boge Brant, Ingrid Glad, Ingrid Hobæk Haff, Jens Kristoffer Haug, Inge Helland, Aliaksandr Hubin, Jonas Moss, Jonathan Poyti Stang, Vegard Stikbakke, Martin Tveten.

From elsewhere at the University of Oslo: Andrea Cremaschi, Solveig Engebretsen, Arnoldo Frigessi, Torunn Heggland, T.Tien Mai, Marissa LeBlanc, Leiv Tore Salte Rønneberg, Tore Schweder, David Swanson, Magne Thoresen, Ying Yao, Zhi Zhao, Manuela Zucknick.

From elsewhere: Daumantas Bloznelis (Inland Norway University of Applied Sciences), Marta Crispino (INRIA Grenoble), Gerda Claeskens (KU Leuven), Idris Eckley (Lancaster University), Federico Ferraccioli (Università degli studi di Padova), Håkon Gjessing (Norwegian Institute of Public Health), Jan Hannig (University of North Carolina at Chapel Hill), Martin Jullum (Norwegian Computing Centre), Per Mykland (University of Chicago), Håvard Nygård (Peace Research Institute Oslo), Jonas Christoffer Lindstrøm (Akershus University Hospital), Sylvia Richardson (University of Cambridge), Bjarne Røsjø (Titan), Yngve Vogt (Apollon), Alex Whitmore (McGill University), Lan Zhang (University of Illinois at Chicago).

A crowd of eager participants at Ingeniørenes Hus.