Building Bridges (at Bislett)

International FocuStat Workshop:
Bridging Parametrics and Nonparametrics

May 22–24 2017 **Teknologihuset, Oslo**

focustat
FOCUS DRIVEN STATISTICAL
INFERENCE WITH COMPLEX DATA

UiO : University of Oslo

The Research Council of Norway

Building Bridges Workshop
Oslo, May 22–24, 2017

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 followed strictly.

Monday:

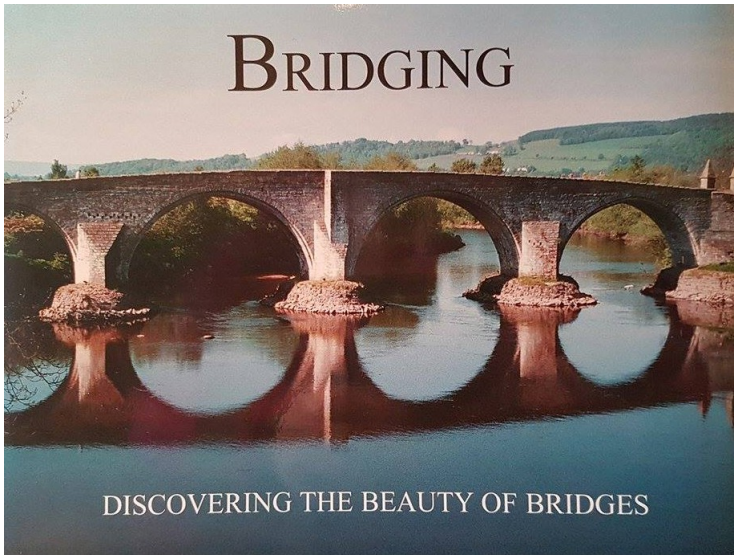
8:45 - 9:10	Good morning & tea and coffee
9:15 - 9:40	**Nils Lid Hjort**: Welcoming remarks: FocuStat; general workshop themes; looking ahead
9:45 - 10:10	**Ingrid Glad:** Old bridge still standing – nonparametric estimation with a parametric start *tea & coffee*
10:30 - 11:15	**Dag Tjøstheim:** Local Gaussian approximation
11:20 - 12:00	**Bård Støve:** Applications of local Gaussian correlation in finance *lunch*
13:30 - 14:00	**Håkon Otneim:** Multivariate density estimation using the local Gaussian correlation *tea & coffee*
14:15 - 14:40	**Céline Cunen:** Confidence distributions: a potential bridge for combining parametric and non-parametric analyses
14:45 - 15:15	**Jonas Moss:** Nonparametric density estimation on $[0, 1]$ with Bernstein splines

Tuesday:

9:00 - 9:25	**Emil Aas Stoltenberg:** Inference in misspecified cure models
9:30 - 10:15	**Sonia Petrone:** On Bayesian nonparametric regression
10:20 - 11:00	**Igor Prünster:** Does the full support property suffice for Bayesian nonparametrics? *tea & coffee*
11:20 - 11:55	**Nils Lid Hjort:** Minimum dispair, maximum despair *lunch*
13:30 - 14:10	**Martin Jullum:** A focused model selection criterion for selecting among parametric and nonparametric models
14:15 - 14:45	**Sam-Erik Walker:** Focused model selection and inference using robust estimators *tea & coffee*
15:00 - 15:25	**Vinnie Ko:** Focused information criteria for copulae
17:30 -	Mini-excursion & Workshop Dinner & Applied Bridgeology

Wednesday:

9:00 - 9:45	**Thore Egeland:** Parameters and pedigrees
9:45 - 10:30	**Christian Rohrbeck:** Bayesian spatial monotone multiple regression *tea & coffee*
10:45 - 11:15	**Nils Lid Hjort:** The hybrid likelihood: Combining parametrics and nonparametrics
11:15 - 12:00	**Riccardo de Bin:** Integrated likelihoods in the presence of many nuisance parameters *lunch*
13:30 - 14:15	**Gudmund Hermansen:** Focused regularised likelihood
14:15 - 14:50	**Kristoffer Hellton:** Focused Fine-tuning of Ridge Regression for Personalized Prediction *tea & coffee*
15:05 - 15:30	Summing up & rounding off



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 project group consists of Nils Lid Hjort (professor, project leader), Gudmund Hermansen and Kristoffer Hellton (PostDocs), 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, 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 will also involve 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

`www.mn.uio.no/math/english/research/projects/focustat/index.html`

and check (and contribute to) our Facebook page.

FocuStat has budget for arranging small-scale international workshops on designated themes, for each of 2015, 2016, 2017, 2018, as well as for lower-scale research kitchens (which have been dedicated to minimum divergence and scoring methods, and to robustification of likelihoods).

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, etc. A Special Issue of *Journal of Statistical Planning and Inference* is being published in the course of 2017, partly based on invited talks to the 2015 workshop.

The **May 2016** workshop was on **FICology**, concerning focused ways in which to build and select models, along with model averaging, post-selection issues, etc. Methods and applications involve uses and variations of Focused Information Criteria.

The present **May 2017** workshop is on **Building Bridges**, broadly speaking meant to invent, assess, examine and fruitfully utilise 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.

A bigger conference is planned for **May 2018**, where themes of earlier workshops are revisited, along with new types of applications.

Titles & abstracts:

Riccardo de Bin:

Integrated likelihoods in the presence of many nuisance parameters

Abstract: Frequentist inference about a parameter of interest in presence of a nuisance parameter can be based on an integrated likelihood function. We analyze the behaviour of inferential quantities based on such a pseudo-likelihood in a two-index-asymptotics framework, in which both sample size and the number of nuisance parameters may diverge to infinity. We show that a properly chosen integrated likelihood largely outperforms standard likelihood methods, such as those based on the profile likelihood. These results are confirmed by simulation studies, in which comparisons with modified profile likelihood are also considered.

Céline Cunen:

Confidence distributions: a potential bridge for combining parametric and nonparametric analyses

Abstract: Combining information across different sources is an important statistical challenge, arising in many different fields. Combination is especially difficult when the sources are very diverse, requiring approaches beyond standard meta-analysis methods. The II-CC-FF paradigm (Independent Inspection, Confidence Conversion, Focused Fusion) is a general three-step method for such problems. The first step, II, uses different techniques to translate the information from each source to confidence distributions. Then these confidence distributions are transformed into confidence log-likelihoods in the CC-step, before being combined in the FF step. In this talk, the II-CC-FF scheme will be presented, and I will indicate how it may be used in order to combine inference about a common parameter when some of the sources have been analysed by parametric and others via nonparametric methods. Some examples will be given, and other confidence-distribution-based combination methods will also be briefly discussed. (This is partly joint work with Nils Lid Hjort.)

Thore Egeland:

Parameters and pedigrees

Abstract: I will discuss statistical models for the family relationship between members of a pedigree. The simplest situation occurs for two non-inbred individuals: their relationship is described by two parameters. If inbreeding is possible, i.e., the parents are related within the pedigree, 9 parameters, the Jacquard coefficients, characterise the relationship. The number of parameters needed for relationships involving more than two individuals increases rapidly. The extensions beyond the pairwise case will focus on problems rather than solutions as little is known.

Ingrid Glad:

Old bridge still standing – nonparametric estimation with a parametric start

Abstract: Twenty years ago, in Hjort & Glad (1995) and Glad (1998a, 1998b), we introduced an innovative bridge between two worlds using multiplicative parametric 'starts' or 'guides' in nonparametric density and curve estimation. If the parametric start is not too far from the true underlying model, this semiparametric construction reduces bias without increasing variance when compared to the nonparametric sibling – hence offering some sort of a free lunch. This idea has later been elaborated by others in many different settings, and extended with an additive version as well. Lately it has even turned up in connection to big data and embarrassingly parallel MCMC.

I will remind us about the original idea and theory, and give some examples of recent developments, as inspiration to building new bridges from good, old stones.

Kristoffer Hellton:

Focused Fine-tuning of Ridge Regression for Personalized Prediction

Abstract: Statistical prediction methods typically depend on tuning parameter(s) requiring some form of fine-tuning. For ridge regression, least squares regression with an L2 penalty, there is a range of fine-tuning procedures with K-fold cross-validation as the standard, but common for all is the choice of one tuning parameter for all future predictions. With the explicit solution of ridge regression, one can however focus the tuning parameter to give optimal prediction for a specific set of covariates. Each new outcome to be predicted is thereby given a unique tuning parameter, resulting in a personalized prediction. We introduce the focused ridge – fridge – procedure, where we minimize an estimated mean squared error of the prediction, through plug-in estimates of the regression and error variance parameters. Our procedure is illustrated by two applications; predicting personalized disease risk and treatment response based on (high dimensional) genetic data. Finally, we discuss the possibility of additional covariate selection, on top of the ridge regression model, using FIC. (This is partly joint work with Nils Lid Hjort.)

Gudmund Hermansen:

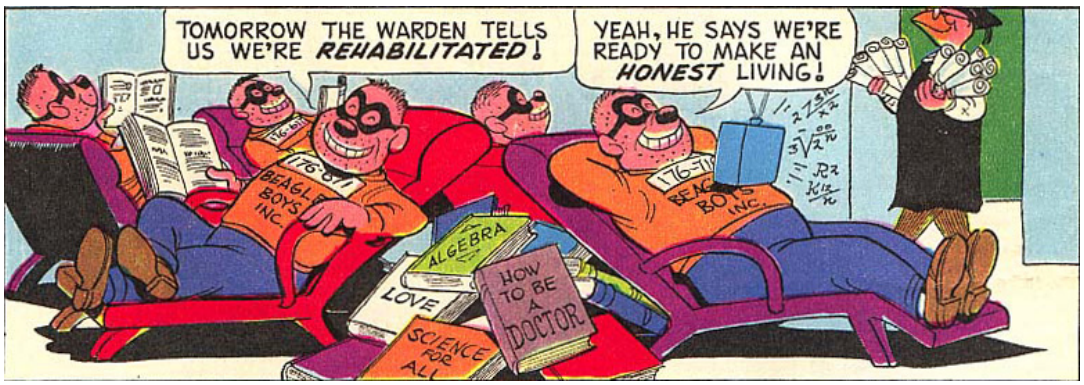
Focused regularised likelihood

Abstract: We introduce a general strategy for focused and robust parameter estimation. The underlying idea is to maximise a penalised version of the log-likelihood $\ell_n(\theta)$, where the penalisation is selected to secure and control that the model behaves in accordance with a predefined focus parameter ψ estimated under an alternative model. The control parameter $\psi(\theta)$ – judged to be of most importance – will be compared with $\widehat{\psi}$ estimated under an alternative model; typically based on a nonparametric procedure. A new estimator is obtained from maximising the focused regularised log-likelihood (FRL) $\ell_n(\theta) - \frac{1}{2}n\lambda(\widehat{\psi} - \psi(\theta))^2$ for a tuning parameter λ. In contrast to other approaches, the regularisation scales with sample size, meaning that the influence of the alternative model will not be washed out. The large-sample properties of the FRL estimator is explored in various models, and we show that our FRL approach is able to beat (in terms of mean squared error for ψ) the maximum likelihood under certain types of model misspecifications. (This is partly joint work with Nils Lid Hjort.)

Nils Lid Hjort:

Opening remarks

Abstract: I will first sketch some of the intended research lines for the FocuStat project (January 2014 to December 2018), pointing also to the *Inference With Confidence* and *FICology* workshops in May 2015 and May 2016. Some remarks will also be made regarding 2018 activities. Then I will attempt to describe some of the themes for the present *Building Bridges (at Bislett)* workshop, including pointers to yet other bridge-building or related themes which also could have been presented.



BBB-gjengen

Nils Lid Hjort:

The hybrid likelihood: Combining parametric and nonparametric likelihoods

Abstract: The FIC approach is in essence to select among candidate strategies by evaluating, assessing and estimating the performance of the different competitors, and then select the winner, the method with best estimated risk. This applies fruitfully to situations where these competing schemes are associated with different models. The general idea can be applied also in other contexts, however, like choosing the best estimator inside a class of estimators, the best confidence distribution across a spectrum of possibilities, etc.

My talk will consist of two parts. First I introduce a method for melding together the classic parametric likelihood with the nonparametric empirical likelihood, and work out theory for such schemes. The method is really a class of methods, as the statistician would need to choose both which extra parameters to include in the construction, and a certain balance parameter that weighs parametrics against nonparametrics. I will attempt to show how FIC thinking may be used to help in these choices. (This is partly joint work with Ian McKeague and Ingrid Van Keilegom.)

Nils Lid Hjort:

Minimum dispair, maximum despair

Abstract: Consider the divergence $d(g, f_\theta) = \int B(g/f_\theta) f_\theta \, \mathrm{d}y$ between densities, where $B(\cdot)$ is smooth, strictly convex, and with $B(1) = 0$. From Jensen, such a distance is nonnegative and zero only when the two densities are equal. The case of $B_1(\rho) = \rho \log \rho$ corresponds to classical Kullback–Leibler; taking $B_2(\rho) = -\log \rho$ corresponds to 'backwards Kullback–Leibler'; with $B_3(\rho) = 1 - \sqrt{\rho}$ one finds the Hellinger distance, etc. A general estimating strategy is to choose $\widehat{\theta}$ to minimise $d(\widetilde{g}, f_\theta)$, where $\widetilde{g}$ is a separate estimate of the data density. A version of this, with $B_1(\cdot)$, leads to the maximum likelihood method. Other estimator versions emerge with other choices of $B(\cdot)$, for suitable pilot estimators $\widetilde{g}$. The so-called minimum disparity parametric estimators correspond to taking a nonparametric $\widetilde{g}$.

I will discuss certain versions of these schemes, pointing also to extensions of methods and results found in the chief textbook on such methods, Basu, Shioya, Park (2011), 'Statistical Inference: The Minimum Distance Approach'. Parts of what I present will be related to ongoing work with Sam-Erik Walker.

Martin Jullum:

A focused model selection criterion for selecting among parametric and nonparametric models

Abstract: A long-living question is whether one should rely on a parametric or nonparametric model when analyzing 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. The underlying principle of the FIC has earlier been utilized to develop model selection criteria in situations involving mainly nested parametric models. The present approach allows general non-nested parametric models to be compared to a nonparametric candidate model. Including an asymptotically unbiased nonparametric candidate model is beneficial as it 'saves the day' when none of the parametric models perform well – while leaving estimation to the parametric models when they do perform well. For presentational simplicity we shall mainly work with the i.i.d. setting, but we shall also discuss extensions to semiparametric vs. fully parametric Cox regression, and to nonparametric vs. parametric time series models. Finally, we take a look at some theoretical properties of the developed criterion.

Vinnie Ko:

FIC for copulae

Abstract: Copula is a flexible way of modelling dependency between variables and it can be easily generalized into high dimensions. There are three most popular estimation strategies for estimating copula parameters and margins: ML (maximum likelihood, parametric), IFM (inference function for margins, stepwise parametric) and PML (pseudo maximum likelihood, semiparametric). We (try to) derive FIC formulae for each situation. A simulated dataset will be used as an example. Choices for focus parameters in the copula setting will be discussed.

Jonas Moss:

Nonparametric density estimation on $[0, 1]$ *with Bernstein splines*

Abstract: Nonparametric density estimation using splines can be done in several ways, for instance by converting histograms to data points. These methods work well, but do not yield bona fide densities, and are prone to boundary bias when used on a compact interval such as [0,1]. More common than spline density estimation is kernel density estimation, where the standard kernels yield highly inefficient estimates when used on $[0, 1]$, mainly due to boundary bias. Some non-standard kernels mitigate the boundary bias problem, but still suffer from high space complexity – a kernel density estimate requires at least one parameter for each point. In this talk I present the new method of Bernstein splines density estimation, an almost parametric method that requires far fewer parameters than kernel density estimators, at the cost of high computational complexity. The method fits the data to a mixture of disjoint Bernstein densities, which subsequently are joined together in the same way as splines.

Håkon Otneim:

Multivariate density estimation using the local Gaussian correlation

Abstract: We take the local Gaussian correlation out from the comforts of bivariate analysis, and try to tackle the problem of multivariate density estimation. The classical kernel density estimator suffers heavlily under the curse of dimensionality, which, unfortunately, is not something that can be solved by fast computers or clever mathematics. It is simply a fact of life that we have to deal with.

What the local Gaussian correlation offers, however, is an intuitive way to break the problem of multivariate density estimation down to a series of bivariate estimation problems. This offers an appealing middle ground where we can obtain reliable estimates of the multivariate density function, while still remaining fairly agnostic about the dependence structure between variables. This has resulted in the locally Gaussian density estimator (LGDE), which performs well in a variety of simulation experiments.

A key insight in this talk is that the local Gaussian correlation relates to a general stochastic vector in a similar way as the ordinary correlation coefficient relates to a jointly normally distributed vector. That means that we can not only use it to test for independence, but also to estimate conditional density functions, and to characterize and test for conditional independence by calculating the partial local Gaussian correlation.

Sonia Petrone:

On Bayesian nonparametric regression

Abstract: The talk aims at giving an overview and discussion of some approaches to Bayesian nonparametric regression. The growth of Bayesian nonparametrics has seen an explosion of models and methods, in the statistics and machine learning communities. The talk will restrict the attention on density regression, or estimation of a conditional density, through Dirichlet mixture models. Even so, the literature is vast and fragmented. We try to give a unifying viewpoint, and address comparison among different proposals.

Our underlying point is that, even beyond frequentist asymptotic properties, fine details of the nonparametric prior may have a relevant impact on the finite sample properties and on the predictive performance. A challenge is to provide a computational strategy that can be fairly easily adapted to the different models under examination. (This is joint work with Sara Wade and Michele Peruzzi.)

Igor Prünster:

Does the full support property suffice for Bayesian nonparametrics?

Abstract: Nonparametric priors are typically required to have a large topological support. This allows to construct a nonparametric envelope around a given parametric model within its support. However, this feature is not always sufficient to guarantee the desired distributional flexibility of functionals of the nonparametric prior, which in turn is needed to achieve accurate posterior inferences. This issue is effectively illustrated within the family of Gibbs-type priors, a recent intuitive generalization of the Dirichlet and the Pitman-Yor process priors. In particular, the distribution of certain functionals of interest in species sampling problems and mixture models are considered. It is shown that the Dirichlet process leads to severe restrictions, unlike other Gibbs-type priors that induce highly flexible behaviours.

Christian Rohrbeck:

Bayesian spatial monotonic multiple regression

Abstract: In several regression problems, it can be assumed that the mean response increases with increasing values of the explanatory variables. Inference under this assumption is termed monotonic regression. This talk considers monotonic, multiple regression for a set of contiguous regions (lattice data). The regression functions permissibly vary between regions and may exhibit spatial structure. New Bayesian nonparametric methodology is developed that allows for both continuous and discontinuous functional shapes and which are estimated using marked point processes and reversible jump Markov chain Monte Carlo techniques. Spatial dependence is incorporated by a flexible prior distribution; the parametrization allows the dependence to vary with functional level. The approach is tuned using Bayesian global optimization and cross-validation. Posterior realizations enable variable selection, threshold detection and prediction as well as extrapolation of the regression function. Performance and flexibility of the approach is illustrated by simulation studies and application to a Norwegian insurance data set. This is joint work with Deborah Costain and Arnoldo Frigessi.

Emil Stoltenberg:

Inference in misspecified cure models

Abstract: A cure model is a type of mixture survival model where only a fraction of the subjects under study are deemed susceptible to the event of interest, the remaining fraction are 'cured' in the sense that they will never experience this event. One part of the model determines whether an individual is cured or not (the incidence part), the other part the time to the event for the non-cured fraction (the latency part). In certain epidemiological and clinical studies interest is only in the incidence part of the model, however, statistical inference on the incidence part is dependent on the modelling of the latency part. In some current work (jointly with Per Mykland) we develop methods for comparing modelling strategies for the latency part (the survival times) of a cure model based on the the models' performance in providing precise inference for the incidence part.

Bård Støve:

Applications of Local Gaussian Correlation in Finance

Abstract: A number of studies have provided evidence that financial returns exhibit asymmetric dependence, such as increased dependence during bear markets, but there seems to be no agreement as to how such asymmetries should be measured. In this talk we will use the Local Gaussian Correlation (LGC) to study this asymmetry, and also point at applications, for instance how to measure financial contagion. We further provide evidence that the dependence can vary in time. Modelling this time dynamics of the dependence may call for a parametric version of the LGC, and we provide some ideas in this direction.

Dag Tjøstheim:

Local Gaussian approximation

Abstract: The principle of local Gaussian approximation consists in approximating a multivariate density locally by a multivariate Gaussian density. This enables one to define a local correlation matrix and local means. An overview of recent theory and applications of this principle will be given. The theory is based on local likelihood estimation of locally parametric models, but specialized to a local Gaussian approximant. The theory is not confined to the iid case. In the time series case this leads to a concept of local autocorrelation and local spectral analysis. This type of local analysis can detect nonlinear features which are difficult to detect by global analysis. Applications and perspectives on copula identification, principal components and discrimination analysis will also be briefly mentioned.

Sam-Erik Walker:

Focused model selection and inference using robust estimators

Abstract: Selection of parametric models based on general information criteria like Akaike's information criterion (AIC), Bayesian information criterion (BIC), and similar, is a well-established practice within the statistical science. Over the years, more specific model selection criteria have also been developed, like the so-called focused information criterion, or FIC, where models are selected based on specifying a focus parameter – a certain parameter or function of the data distribution deemed most important in a statistical setting. Up to now FIC has been mostly based on the maximum likelihood estimator, which is efficient, but not robust. In this presentation, we indicate a possible extension of the theory of the FIC to the use of robust estimators, with a nonparametric robust alternative, to perform focused model selection and inference in situations where some of the data might be contaminated with atypical or erroneous values, outliers etc. Under appropriate conditions this will secure a certain higher degree of robustness of the final methods. Some examples will illustrate the new methodology. (This is partly joint work with Nils Lid Hjort.)

Practical information:

Most of our workshop guests will stay at Scandic Holberg, a 700 meters walk from Teknologihuset where the Workshop will take place (Pilestredet 56). If one prefers to use public transportation, one can take the trams 17 or 18 (direction 'Rikshospitalet') from the stop 'Holbergs plass' to 'Dalsbergstien'.

Both Teknologihuset 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 Olympic stadium Bislett is a javelin's throw away from Teknologihuset (see the cover photo by Paul A. Røstad showing Ivar Bredholt setting a new personal record of 4.30 m, precisely fifty years ago, in the pre-catapult era of things), and 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 Bjørnsen, Henrik Wergeland, etc. If one wishes to admire some art by Munch, the National Gallery is found only 300 meters from the hotel.

Map of area between Scandic Holberg and Teknologihuset. Blue lines are Tram lines. *Google maps*

BBB Dinner:

On Tuesday the 23rd, the FocuStat group invites all participants to join us for a social, cultural and gastronomical evening.

We will walk together from 'Holbergs plass' at approximately 17:30. At 18:00 we will have a guided tour of *Det astronomiske observatorium* (The Observatory), a former astronomical research station and the oldest building of the University of Oslo. Then, at 19:30, we will dine at the nearby restaurant 'BA53', a stylish neighbourhood brasserie. The restaurant specialises in modern cuisine using Nordic ingredients, with a playful and elegant twist.

The Astronomical Observatory (*Photo by Lars Mæhlum*)

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: Riccardo de Bin, Arnoldo Frigessi, Ingrid Glad, Jonas Moss.

From elsewhere: Thore Egeland (Norwegian University of Life Sciences), Martin Jullum (Norwegian Computing Centre), Håkon Otneim (Norwegian School of Economics), Sonia Petrone (Bocconi University), Igor Prünster (Bocconi University), Christian Rohrbeck (Lancaster University), Bård Støve (University of Bergen), Dag Tjøstheim (University of Bergen).