

# Likelihood Inference for Diffusions

Yacine Ait-Sahalia

Bendheim Center for Finance Princeton University, Princeton University

This talk surveys recent results on closed form likelihood expansions for discretely sampled diffusions. One major impediment to both theoretical modeling and empirical work with continuous-time models is the fact that in most cases little can be said about the implications of the instantaneous dynamics for longer time intervals. One cannot in general characterize in closed form an object as simple, yet fundamental for everything from prediction to estimation and derivative pricing, as the conditional density of the process, also known as the transition function of the process. I will describe a method which produces accurate approximations in closed form to the transition function of an arbitrary multivariate diffusion. I will then show a connection between this method and saddlepoint approximations and provide examples. Next, I will discuss inference using this method when the state vector is only partially observed, as in stochastic volatility or term structure models. Finally, I will outline the use of this method in specification testing and sketch derivative pricing applications.

## The Limit of Finite Sample Size and a Problem with Subsampling

Donald W.K. Andrews

Department Economics, Yale University

This paper considers tests and confidence intervals based on a test statistic that has a limit distribution that is discontinuous in a nuisance parameter or the parameter of interest. The paper shows that standard fixed critical value (FCV) tests and subsample tests often have asymptotic size—defined as the limit of the finite sample size—that is greater than the nominal level of the test. We determine precisely the asymptotic size of such tests under a general set of high-level conditions that are relatively easy to verify. Often the asymptotic size is determined by a sequence of parameter values that approach the point of discontinuity of the asymptotic distribution. The problem is not a small sample problem. For every sample size, there can be parameter values for which the test over-rejects the null hypothesis. Analogous results hold for confidence intervals.

We introduce a hybrid subsample/FCV test that alleviates the problem of over-rejection asymptotically and in some cases eliminates it. In addition, we introduce size-corrections to the FCV, subsample, and hybrid tests that eliminate over-rejection asymptotically. In some examples, these size corrections are computationally challenging or intractable. In other examples, they are feasible. This is joint work with Patrik Guggenberger.

# Very High-dimensional Data: Prediction and Variable Selection

Peter Bühlmann

Swiss Federal Institute of Technology Zurich

We consider problems where the number of predictor variables  $p$  is much larger than sample size  $n$ , i.e. function, the Lasso or also boosting algorithms have been shown to be asymptotically consistent and both of them often exhibit very good empirical performance. However, the problem of variable selection is much more subtle and difficult than prediction.

We will discuss theoretical and practical potential and limitations of the Lasso and boosting for variable selection, and we will present powerful improvements. The talk is a special birthday tour for Peter Bickel: from "Relaxed Lasso" over "Sparse Boosting" to completely different ideas from the "PC algorithm" in graphical modeling. The methods are used for two problems in computational biology: (i) alternative splicing using single-gene libraries; and (ii) short motif modeling for splice site detection.

## Powerful Choices: Variable and Tuning Constant Selection in Nonparametric Regression based on Power

Kjell Doksum

Department of Statistics, University of California, Berkeley

This paper considers nonparametric multiple regression procedures for analyzing the relationship between a response variable and a vector of covariates. It uses an approach which handles the dilemma that with high dimensional data the sparsity of data in regions of the sample space makes estimation of nonparametric curves and surfaces virtually impossible. This is accomplished by abandoning the goal of trying to estimate true underlying curves and instead estimating measures of dependence that can determine important relationships between variables. These dependence measures are based on local parametric fits on subsets of the covariate space that vary in both dimension and size within each dimension. The subset which maximizes a signal to noise ratio is chosen. The signal is a local estimate of a dependence parameter which depends on the subset size, and the noise is an estimate of the standard error (SE) of the estimated signal. This approach of choosing the window size to maximize a signal to noise ratio lifts the curse of dimensionality because for regions with sparsity of data the SE is very large. For contiguous Pitman alternatives it corresponds to asymptotically maximizing the probability of correctly finding relationships between covariates and a response, that is, maximizing asymptotic power. It is shown that within a selected dimension, the bandwidths of the optimally selected subset

do not tend to zero as the sample size  $n$  grows except for alternatives where the length of the intervals where the alternative differs from the hypothesis tends to zero as  $n$  grows. One of the dimension reduction algorithms is used together with MARS and GUIDE and is shown to improve their performance. This is joint work with Chad Schafer, Shijie Tang and Kam Tsui.

## **Sparsity in Inference: Past Trends, Future Promise**

David Donoho

Statistics Department, Stanford University

Suppose we have to estimate a large number of parameters, most of which are zero or negligible and some of which are important or significant; but we don't know in advance which parameters are likely to be negligible and which are likely to be important. This important problem in some sense spans large swaths of applied statistics, from regression model building to gene association studies.

I'll discuss some of Peter Bickel's early work related to this problem, and how the problem has grown and mutated over the years. At this point, it's a problem with truly vast implications, having applications throughout science and technology, with lots of challenging mathematics and surprising applications.

## **Methods of Robust Online Signal Extraction and Applications**

Ursula Gather

Department of Statistics, University of Dortmund

We discuss filtering procedures for robust extraction of a signal from noisy time series. These methods can e.g. be applied to online observations of vital parameters which are acquired by clinical information systems for critically ill patients. Multivariate time series from online monitoring exhibit trends, abrupt level changes and large spikes (outliers) as well as periods of relative stability. Also, the measurements are overlaid with a high level of noise and among the variables strong dynamic dependencies are found (Gather et al. (2002)). The challenge is to develop methods that allow a fast and reliable denoising of these time series. Noise and artifacts are to be separated from structural patterns of relevance.

Standard approaches to univariate signal extraction are moving averages and (univariate) running medians, but they have shortcomings when outliers or trends occur. Reviewing and extending recent work we present new methods for robust online signal extraction and discuss their merits for preserving trends, abrupt shifts and extremes and for the removal of spikes (Davies, Fried, Gather (2004)). Our robust regression moving window

methods are applicable even in real time because of increased computational power and fast algorithms (Bernholt and Fried (2003)).

In multivariate robust signal extraction efficiency is lost if the error terms of the variables are highly correlated since generalizing robust univariate regression methods does not result in affine equivariant procedures. Multivariate affine equivariant regression methods with high breakdown, as e. g. MCD-regression (Rousseeuw et al. (2004)), more over assume that the data are in general position. For discrete data in short time windows this is however often not the case.

We therefore propose new procedures for multivariate signal extraction, which offer fast and robust signal extraction, good efficiency properties and which can be used for discretely measured data with low variability as well as in situations with many outliers.

## **Convergence and Consistency of Newton's Algorithm for Estimating a Mixing Distribution**

Jayanta K. Ghosh

Department of Statistics, Purdue University

In recent years Michael Newton has proposed an algorithmic estimate of a mixing distribution, which is computationally efficient. We prove its convergence and consistency under rather strong conditions. The consistency result is new. A proof of convergence given earlier under same conditions by Newton is shown to be incomplete and not easily rectifiable. We study various other aspects of the estimate and compare it with the Bayes estimate based on Dirichlet mixtures. This is joint work with Surya Tokdar.

## **Edgeworth Approximations for Symmetric Statistics**

Friedrich Goetze

Department of Mathematics, University of Bielefeld

We shall describe conditions, such that Edgeworth approximations up to an error  $o(N^{-1})$  hold for a general class of asymptotical linear symmetric statistics in  $N$  independent observations, which admits a regular stochastic Hoeffding expansion. The conditions involve Cramer's condition of smoothness for the linear term and some covariance type conditions for the second order term. The results are joint work with M. Bloznelis and extend previous work by P. Bickel, V. Bentkus, W. van Zwet and the author. They are based on new analytical and combinatorial techniques. Connections with approximation results in Probability and Number Theory for related degenerate  $U$ -statistics, and their dimension dependence will be discussed as well.

# Some Theory for Classifiers in High-dimensional, Low Sample Size Settings

Peter Hall

Centre for Mathematics and its Applications, Mathematical Sciences Institute,  
Australian National University

A large class of distance-based classifiers is defined, and their performance addressed using theoretical arguments based on letting dimension diverge as sample size is kept fixed. Particular attention is paid to the use of truncation, to heighten sensitivity of the classifiers in cases of data sparsity. It is shown that in that setting, truncated distance-based classifiers can perform well when differences between distributions are detectable but not estimable. They do not do quite as well as classifiers based on Donoho and Jin's higher-criticism methods, although they are more robust against assumptions about distribution type and component relationships. However, the robustness of higher criticism can be increased by using methods based thresholding, as well as empirical approaches.

## A Statistical Framework to Infer Functional Gene Associations from Multiple Biologically Dependent Microarray Experiments

Haiyan Huang

Department of Statistics, University of California, Berkeley

Microarray data from an increasing number of biologically interrelated and interdependent experiments now allow more complete portrayals of functional gene relationships involved in biological processes. However, in the current integrative analyses of microarray data, an important practical issue is widely ignored: the existence of dependencies among gene expressions across biologically related experiments. When not accounted for, these dependencies (due to either similar intrinsic conditions or relevant external perturbations among the experiments) can result in inaccurate inferences of functional gene associations, and hence incorrect biological conclusions. To address this fundamental problem, we propose a new measure,  $K_{\text{norm}}$  correlation, to quantify functional gene associations in the presence of such experimental dependencies. Our intuitive strategy is to reduce the experimental dependencies before estimating gene correlations. The statistical model underlying  $K_{\text{norm}}$  correlation is a multivariate normal distribution characterized by a Kronecker product dependency structure. This unique structure maintains the same experimental correlations across genes and the same gene correlations across experiments. The proposed measure simplifies to the Pearson coefficient when experiments are uncorrelated. Applications to simulation studies and to two real datasets (on yeast and human

genes) demonstrate the success of Knorm correlation, and also the adverse impact of experimental dependencies on gene associations using Pearson coefficients. Knorm correlation is expected to greatly improve the accuracy of biological inferences made from experiments currently (and incorrectly) assumed to be uncorrelated.

This is a joint work with Melinda Teng and Xianghong Zhou.

## **Fence Methods: Another Look at Model Selection**

Jiming Jiang

Department of Statistics, University of California, Davis

Many model search strategies involve trading off model fit with model complexity in a penalized goodness of fit measure. Asymptotic properties for these types of procedures in settings like linear regression and ARMA time series have been studied. Yet, such strategies do not always translate into good finite sample performance. The issue is typically one of the procedure being overly sensitive to the setting of penalty parameters, which are required to be increasing functions of sample size. Furthermore, these strategies do not generalize naturally to more complex models, such as those for modeling clustered data or those that involve adaptive estimation. In these cases, penalties and model complexity may not be naturally defined.

We introduce a new class of model selection strategies known as fence methods. The general idea involves a procedure to isolate a subgroup of what are known as correct models (of which the optimal model is a member). This is accomplished by constructing a statistical fence, or barrier, to carefully eliminate incorrect models. Once the fence is constructed, the optimal model will be selected among the correct models (those within the fence) according to simplicity of the models. We describe a variety of fence methods, based on the same principle but applied to different situations. These include regression, least angle regression, linear mixed models for clustered and non-clustered data, generalized linear mixed models for clustered and non-clustered data, and time series models. We show the broad applicability of fence methods to all of these areas by giving a number of examples, each supported by simulation results or real-life data analyses. In terms of theoretical development, we give sufficient conditions for consistency of fence, a desirable property for a good model selection procedure.

This work is joint with J. Sunil Rao, Zhonghua Gu and Thuan Nguyen.

# Goodness-of-fit Testing in Interval Censoring Case 1

Hira L. Koul

Department of Statistics and Probability, Michigan State University

In the interval censoring case 1, an event occurrence time is unobservable, but one observes an inspection time and whether the event has occurred prior to this time or not. The focus here is to provide tests of goodness-of-fit hypothesis pertaining to the distribution of the event occurrence time. The proposed tests are based on certain marked empirical processes for testing a simple hypothesis and their martingale transforms. These tests are asymptotically distribution-free, consistent against a large class of fixed alternatives and have nontrivial asymptotic power against a large class of local alternatives.

## Edgeworth Expansions for Sums of Block-variables under Weak Dependence

Soumendra N. Lahiri

Department of Statistics, Iowa State University

Let  $\{X_i\}_{i=-\infty}^{\infty}$  be a sequence of random vectors and let  $Y_{in} = f_{in}(\mathcal{X}_{i,\ell})$  be zero mean *block-variables* where  $\mathcal{X}_{i,\ell} = (X_i, \dots, X_{i+\ell-1})$ ,  $i \geq 1$  are overlapping blocks of length  $\ell$  and where  $f_{in}$  are Borel measurable functions. This paper establishes valid joint asymptotic expansions of general orders for the joint distribution of the sums  $\sum_{i=1}^n X_i$  and  $\sum_{i=1}^n Y_{in}$  under weak dependence conditions on the sequence  $\{X_i\}_{i=-\infty}^{\infty}$  when the block length  $\ell$  grows to infinity. In contrast to the classical Edgeworth expansion results where the terms in the expansions are given by powers of  $n^{-1/2}$ , the expansions derived here are mixtures of two series, one in powers of  $n^{-1/2}$  while the other in powers of  $[\frac{n}{\ell}]^{-1/2}$ . Applications of the expansions to studentized statistics and to block bootstrap methods for time series data are given.

## Detection in Wireless Sensor Networks

Elizaveta Levina

Department of Statistics, The University of Michigan

Wireless sensor networks are becoming more widely available for use in various applications, such as intruder detection and ecological monitoring. The basic issues in sensor networks (detection, estimation, design) are statistical but little work in this area has been

done by statisticians. I will give a brief overview of the main problems and then focus on a local-vote decision algorithm we developed for target detection by a wireless sensor network. Sensors acquire measurements corrupted by noise, make individual decisions, correct their decisions after consulting the neighboring sensors, and then a collective decision is made by the network. Related local methods have been proposed by the engineers but no theoretical performance guarantees were available. We give an explicit formula for the decision threshold for a given false alarm rate, based on limit theorems for weakly dependent random fields. We also show that, for a fixed false alarm rate, the local-vote correction significantly improves target detection rate.

Joint work with George Michailidis and Natallia Katenka.

## **Bayesian Methods in Haplotype Inference and Disease Mapping**

Jun Liu

Department of Statistics, Harvard University

Haplotypes provide complete information of inheritance, which are very useful in population genetics and association studies. Since experimentally determining haplotype data is expensive, much effort has been devoted to develop computational tools for inferring haplotypes from genotype data. I will present a few Bayesian and semi-Bayesian models that have been formulated over the past few years for this task, including new hierarchical Bayes model developed in our group that incorporates the coalescence effect in a prior distribution. The prediction accuracy of the new method is uniformly improved compared to existing methods such as HAPLOTYER and PHASE.

I will further discuss a Bayesian approach in detecting multi-locus interactions (epistasis) for case-control association studies. Existing methods are either of low power or computationally infeasible when facing of a large number of markers. Using MCMC sampling techniques, the method can efficiently detect interactions among thousands of markers. Using simulation results, I will discuss the power of our approach and the importance to consider epistasis in association mapping.

Based on joint work with Yu Zhang and Tim Niu.

# Mining Massive Text Data: Classification, Construction of Tracking Statistics and Inference under Misclassification

Regina Liu

Department of Statistics, Rutgers University

We present a systematic data mining procedure for exploring large free-style text datasets to discover useful features and develop tracking statistics (often referred to as performance measures or risk indicators). The procedure includes text classification, construction of tracking statistics, inference under error measurements and risk analysis. The main difficulty in deriving this inference scheme is the accounting for misclassification errors, for which we propose two types of approaches: “plug-in” and “projection” methods. We also consider the bootstrap calibration for fine tuning. Finally, as an illustrative example, the proposed data mining procedure is applied to analyzing an aviation safety report repository from the FAA to show its utility in aviation risk management or general decision-support systems.

Although most illustrations here are drawn from aviation safety data, the proposed data mining procedure applies to many other domains, including, for example, mining free-style medical reports for tracking possible disease outbreaks.

This is joint work with Daniel Jeske, Department of Statistics, UC Riverside.

## Statistical Physics and Statistical Computing: A Critical Link— Estimating Criticality via Perfect Sampling

Xiao-Li Meng

Department of Statistics, Harvard University

This talk is based on the following chapter, jointly written with James Serfling of U.S. Department of Defense, in the volume dedicated to Professor Peter Bickel: “The main purpose of this chapter is to demonstrate the fruitfulness of cross-fertilization between statistical physics and statistical computation, by focusing on the celebrated Swendsen-Wang algorithm for the Ising model and its recent perfect sampling implementation by Mark Huber. In particular, by introducing Hellinger derivative as a measure of instantaneous changes of distributions, we provide probabilistic insight into the algorithm’s critical slowing down at the phase transition point. We show that at or near the phase transition, an infinitesimal change in the temperature parameter of the Ising model causes an astronomical shift in the underlying state distribution. This finding suggests an interesting conjecture linking the critical slowing down in coupling time with the grave instability of the system as characterized by the Hellinger derivative (or equivalently, by Fisher information). It also suggests that we can approximate the critical point of the Ising model, a

physics quantity, by monitoring the coupling time of Huber’s bounding chain algorithm, an algorithmic quantity. This finding might provide an alternative way of approximating criticality of thermodynamic systems, which is typically intractable analytically. We also speculate that whether we can turn perfect sampling from a pet pony into a workhorse for general scientific computation may depend critically on how successful we can engage, in its development, researchers from statistical physics and related scientific fields.”

## Smoothing Large Tables

Stephan Morgenthaler

EPFL Learning Center

Methods to smooth large tables are described. Such smoothing problems are of interest in many scientific contexts and with a variety of objectives in mind. One may want to interpolate the table entries, or to quantify the differences between rows and columns, or to classify rows and columns into homogeneous subgroups, or to find the best rows and columns, or some other objective. Fisher’s ANOVA, which can be computed by sweeping row means and column means from the table, assigns a single effect to each row and each column and was originally invented for tables of low dimension. The singular value decomposition of the table offers an alternative single effects approximation. In both cases, the smoothed row traces, that is the plot of the row entries against the row effects, are straight lines.

More general table smoothers are obtained by using more flexible traces. Some of the difficulties with this approach are discussed, among them the choice of row and column variables replacing the single effects from above, the parsimonious choice of trace parameters, the classification of traces, and the transformation of table entries.

## Functional Variance

Hans-Georg Müller

Department of Statistics, University of California, Davis

Functional data consist of an observed sample of smooth random trajectories. A key tool for the analysis of such data is a representation in terms of eigenfunctions of the autocovariance operator of the underlying stochastic process and the associated functional principal components. In some applications the information of interest resides not in the observed smooth random trajectories themselves but rather in the additive noise. Assuming the noise is composed of a white noise component and a smooth random process component, we refer to the latter as the functional variance process. This process can

then be decomposed in terms of its eigenfunctions. Methods to estimate eigenfunctions and functional principal component scores for the functional variance process are based on residuals obtained in an initial smoothing step, applied to the original data. We discuss asymptotic justifications and applications. (joint work with U. Stadtmuller and F. Yao).

## **Statistical Inverse Problems in Active Network Tomography**

Vijay Nair

Department of Statistics, Department of Industrial & Operations Engineering,  
University of Michigan, Ann Arbor

The term network tomography, first introduced in Vardi (1996), characterizes two classes of large-scale inverse problems that arise in the modeling and analysis of computer and communications networks. This talk will deal with active network tomography where the goal is to recover link-level quality of service parameters, such as packet loss rates and delay distributions, from end-to-end path-level measurements. Internet service providers use this to characterize network performance and to monitor service quality. We will provide a review of recent developments, including the design of probing experiments, inference for loss rates and delay distributions, and applications to network monitoring. This is joint work with George Michailidis, Earl Lawrence, Bowei Xi, and Xiaodong Yang.

## **Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration**

Byeong Park

Department of Statistics, Seoul National University

We propose marginal integration estimation and testing methods for the coefficients of varying coefficient multivariate regression model. Asymptotic distribution theory is developed for the estimation method which enjoys the same rate of convergence as univariate function estimation. For the test statistic, asymptotic normal theory is established. These theoretical results are derived under the fairly general conditions of absolute regularity ( $\beta$ -mixing). Application of the test procedure to the West German real GNP data reveals that a partially linear varying coefficient model is best parsimonious in fitting the data dynamics, a fact that is also confirmed with residual diagnostics.

# **Applied Asymptotics**

Nancy Reid

Department of Statistics, University of Toronto

The theory of higher order asymptotics provides quite accurate approximations for a large number of parametric models. However, the details of the theory are somewhat complicated, and perhaps for that reason the methods are not used as often as they might be. I will outline some 'case studies' where improved approximation is readily implemented and illustrate the effects on the resulting inference. I will suggest areas where further research is needed.

## **Multiple Testing in Astronomy**

John Rice

Department of Statistics, University of California, Berkeley

Suppose that a very large number of independent null hypotheses are tested, almost all of which are true. How can the proportion of false null hypotheses be estimated? For motivation, I will discuss the Taiwanese-American Occultation Survey, and will explain how this question arises. I will then present some recent results.

## **Some Remarks on Non-linear Dimension Reduction**

Ya'acov Ritov

Israel Social Sciences Data Center

We remark on the possibility of a well defined dimension reduction. We consider a model in which the data is distributed on a manifold. We present an algorithm for generating a global map of data to a lower dimensional space, minimizing the local structure of the manifold. We remark on the importance on estimating the manifold structure when the main concern is estimating a regression function.

## **Efficient Estimators for Times Series**

Anton Schick

Department of Mathematical Sciences, Binghamton University

I illustrate several recent results on efficient estimation for semiparametric time series models with a simple class of models: first-order nonlinear autoregression with independent innovations. In particular I consider estimation of the autoregression parameter, the innovation distribution, conditional expectations, the stationary distribution, the stationary density, and higher-order transition densities.

# Bayesian Inference in Central Banks: Recent Developments in Monetary Policy Modeling

Christopher A. Sims

Department of Economics, Princeton University

In the 1950's and 60's, large-scale econometric models, grounded in an elegant theory of inference initiated by Trygve Haavelmo, began to be widely used by policy-making institutions. While the models remained in use, their grounding in a theory of inference almost completely disappeared by 2000. In the last few years, there has been research activity in many central banks aimed at producing models grounded in a Bayesian approach to inference and using modern computational approaches to posterior simulation. This talk summarizes the history and describes the methods and results driving the current research.

## Invariant Coordinate Selection (ICS): A Robust Statistical Perspective on Independent Component Analysis (ICA)

David E. Tyler

Department of Statistics, Rutgers University

In many disciplines, *independent component analysis* (ICA) has become a popular method for analyzing multivariate data. Independent component analysis typically assumes the observed data  $\mathbf{Y} \in \mathbb{R}^p$  is generated by a nonsingular affine transformation of independent components, i.e.  $\mathbf{Y} = \mathbf{AZ}$ , where  $\mathbf{A}$  is a nonsingular matrix and  $\mathbf{Z} = (Z_1, \dots, Z_p)'$  consists of independent variables  $Z_1, \dots, Z_p$ . The objective is to then estimate  $\mathbf{A}$  and hence recover  $\mathbf{Z}$ . Approaches for recovering  $\mathbf{Z}$  have often been successful in exploring multivariate data in general, i.e. in cases where the ICA model may not hold. The purpose of this talk is to provide some understanding as to why independent component analysis may work well as a general multivariate method. In particular, without reference to the ICA model, it can be noted that for some methods the recovered  $\mathbf{Z}$  can be viewed as affine invariant coordinates. That is, if we transform  $\mathbf{Y} \rightarrow \mathbf{Y}^* = \mathbf{BY} + \mathbf{b}$  for any nonsingular  $\mathbf{B}$ , then  $\mathbf{Z}^* = \mathbf{\Delta Z} + \mathbf{c}$ , where  $\mathbf{\Delta}$  is a nonsingular diagonal matrix. In other words, the standardized versions of the components  $Z_j$  and  $Z_j^*$  are the same. Hence, the terminology *invariant coordinate selection* (ICS).

Consequently, this leads to the development of a wide class of *affine equivariant coordinatewise methods* for multivariate data. Some methods to be discussed are affine equivariant principal components, robust estimates of multivariate location and scatter, affine invariant multivariate nonparametric tests, affine invariant multivariate distribution functions, and affine invariant coordinate plots. The affine equivariant principal compo-

nents and the corresponding affine invariant coordinate plots can be regarded in a sense as *projection pursuit without the pursuit*. Several examples are given to illustrate the utility of the proposed methods.

## Oracle Inequalities for the LASSO

Sara van de Geer

Seminar fuer Statistik, ETH Zuerich

We consider the LASSO penalty for general M-estimators. Examples include logistic regression, quantile regression, log-density estimation, and boosting with for example logistic loss or hinge loss. Let  $Y$  be a real-valued (response) variable and  $X$  be a (co-)variable with values in some space  $\mathcal{X}$ . Let

$$\mathcal{F} \subset \{f_\alpha = \sum_{k=1}^m \alpha_k \psi_k\}$$

be a (convex subset of a) linear space of functions on  $\mathcal{X}$ . Here,  $\{\psi_k\}_{k=1}^m$  is a given system of base functions. Let  $\gamma_f : \mathbf{R} \times \mathcal{X} \rightarrow \mathbf{R}$  be some loss function, and let  $\{(Y_i, X_i)\}_{i=1}^n$  be i.i.d. copies of  $(X, Y)$ . We consider the estimator

$$\hat{f} = \arg \min_{f_\alpha \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma_f(Y_i, X_i) + \hat{\lambda} \hat{I}(\alpha) \right\},$$

where  $\hat{I}(\alpha) := \sum_{k=1}^m \hat{\tau}_k |\alpha_k|$  denotes the weighted  $\ell_1$  norm of the vector  $\alpha \in \mathbf{R}^m$  with random weights  $\hat{\tau}_k := (\frac{1}{n} \sum_{i=1}^n \psi_k^2(X_i))^{1/2}$ . We study the situation where the number of parameters  $m$  is large (possibly much larger than the number of observations  $n$ ). Our purpose is threefold. Firstly, we want to show that for a proper choice of the smoothing parameter  $\hat{\lambda}$  (possibly depending on  $\{\hat{\tau}_k\}$ ), the estimator  $\hat{f}$  satisfies an oracle inequality. Secondly, we want the result to hold without any a priori bounds on the functions in  $\mathcal{F}$ . Thirdly, we aim at “reasonable” values for the constants involved, as indication that the result is not merely an asymptotic one. In certain settings, the smoothing parameter  $\hat{\lambda}$  can be chosen asymptotically equal to  $4\sqrt{2 \log m/n}$ , which is four times as large as in the linear Gaussian case with soft thresholding. The factor 4 comes from using a symmetrization and a contraction inequality.

# Estimating Function Based Cross-validation

Mark van der Laan

Division of Biostatistics, University of California, Berkeley

Suppose that we observe a sample of independent and identically distributed realizations of a random variable. Given a model for the data generating distribution, assume that the parameter of interest can be characterized as the parameter value which makes the population mean of a possibly infinite dimensional estimating function equal to zero. Given a collection of candidate estimators of this parameter, and specification of the vector estimating function, we propose a norm of the cross-validated estimating equation as criteria for selecting among these estimators. For example, if we use the Euclidean norm, then our criteria is defined as the Euclidean norm of the empirical mean over the validation sample of the estimating function at the candidate estimator based on the training sample. We establish a finite sample inequality of this method relative to an oracle selector, and illustrate it with some examples. This finite sample inequality provides us also with asymptotic equivalence of the selector with the oracle selector under general conditions. We also study the performance of this method in the case that the parameter of interest itself is pathwise differentiable (and thus, in principle, root- $n$  estimable).

## An Expansion for A Discrete Non-lattice Distribution

Willem R. van Zwet

Department of Statistics, University of Leiden

Much is known about asymptotic expansions for asymptotically normal distributions if these distributions are either absolutely continuous or pure lattice distributions. In this paper we begin an investigation of the discrete but non-lattice case. We tackle one of the simplest examples imaginable and find that curious phenomena occur. Clearly more work is needed. (Co-author Friedrich Götze)

## Flexible Approaches to Model Survival and Longitudinal Data Jointly

Jimin Ding and Jane-Ling Wang (Speaker)

Department of Statistics, University of California at Davis

In clinical studies, longitudinal covariates are often used to monitor the progression of the disease as well as survival time. Relationship between a failure time process and some longitudinal covariates is of key interest and so is the understanding of the pattern

of longitudinal process to learn more about health status of patients, or to get some insight into the progression of disease. Joint modeling of the longitudinal and survival data has certain advantages and emerged as an effective way to handle both types of data simultaneously. In this talk, we will explore several intriguing and challenging issues in joint modelling.

Typically, a parametric longitudinal model is assumed to facilitate the likelihood approach. However, the choice of a proper parametric model turns out more illusive than standard longitudinal studies where no survival end-point occurs. Furthermore, the computational burden due to both Monte Carlo numerical integration and EM (Expected Maximum) algorithm is an important concern in the joint modelling setting. To deal with those challenges, we propose several flexible longitudinal models in the joint modelling setting. Simplicity of the model structure is crucial to have good numerical stability, and we will illustrate this through numerical studies and data analysis.

## Goodness of Fit via Phi-divergences: A New family of Test Statistics

Jon A. Wellner

Department of Statistics, University of Washington

A new family of goodness-of-fit tests based on phi-divergences is introduced and studied. The new family is based on phi-divergences somewhat analogously to the phi-divergence tests for multinomial distributions introduced by Cressie and Read (1984), and is indexed by a real parameter  $s \in \mathbf{R}$ :  $s = 2$  gives the Anderson - Darling test statistic,  $s = 1$  gives the Berk-Jones test statistic,  $s = 1/2$  gives a new (Hellinger - distance type) statistic,  $s = 0$  corresponds to the “reversed Berk-Jones” statistic, and  $s = -1$  gives a “studentized” (or empirically weighted) version of the Anderson - Darling statistic. We also introduce corresponding integral versions of the new statistics.

We show that the asymptotic null distribution theory of Jaeschke (1979) and Eicker (1979) for the Anderson-Darling statistic, and of Berk and Jones (1979) applies to the whole family of statistics  $S_n(s)$  with  $s \in [-1, 2]$ . We also provide new finite-sample approximations to the null distributions and show how the new approximations can be used to obtain accurate computation of quantiles.

On the side of power behavior, we show that for  $0 < s < 1$  and fixed alternatives the test statistics always converge almost surely to their corresponding natural parameter. For  $1 < s < \infty$  we provide necessary and sufficient conditions on the alternative d.f.  $F$  for convergence to the corresponding natural parameter to hold, and show that the “Poisson boundary” phenomena noted by Berk and Jones for their statistic continues to hold for  $s \geq 1$  and  $s < 0$  by identifying the Poisson boundary distributions explicitly.

We extend the results of Donoho and Jin (2004) by showing that all our new tests for  $s \in [-1, 2]$  have the same “optimal detection boundary” for normal shift mixture alternatives as Tukey’s “higher-criticism” statistic and the Berk-Jones statistic.

## **Heterogeneous Autoregressive Realized Volatility Model**

Yazhen Wang

Department of Statistics, University of Connecticut

Volatilities of asset returns are pivotal for many issues in financial economics. The availability of high frequency intraday data should allow us to estimate volatility more accurately. Realized volatility is often used to estimate integrated volatility. To obtain better volatility estimation and forecast, some autoregressive structure of realized volatility is proposed in the literature. This talk will present my recent work on heterogeneous autoregressive models of realized volatility.

## **Bayesian Hierarchical Modeling for Integrating Low-accuracy and High-accuracy Experiments**

Jeff Wu

Georgia Institute of Technology, School of Industrial and Systems Engineering

Standard practice in analyzing data from different types of experiments is to treat data from each type separately. By borrowing strength across multiple sources, an integrated analysis can produce better results. Careful adjustments need to be made to incorporate the systematic differences among various experiments. To this end, some Bayesian hierarchical Gaussian process models (BHGP) are proposed. The heterogeneity among different sources is accounted for by performing flexible location and scale adjustments. The approach tends to produce prediction closer to that from the high-accuracy experiment. The Bayesian computations are aided by the use of Markov chain Monte Carlo and Sample Average Approximation algorithms. The proposed method is illustrated with two examples: one with detailed and approximate finite elements simulations for mechanical material design and the other with physical and computer experiments. (Based on joint work with Zhiguang Qian).

# **Semiparametric Mixed Effects Models for Duration and Longitudinal Data**

Zhiliang Ying

Department of Statistics, Columbia University

In this talk, I will present a doubly semiparametric mixed effects model for duration and recurrent event time data. This model is useful in accommodating possible informative censoring, a common problem in many follow-up studies. It also exhibits interesting features which make it relatively easy to carry out the usual statistical inferences. We show the usefulness and practicality of the proposed approach via theoretical properties, simulation results and data analysis. Some additional developments on linear mixed effects model for longitudinal data will also be presented.

# **Spatially Adaptive Functional Linear Regression with Functional Smooth Lasso**

Chunming Zhang

Department of Statistics, University of Wisconsin

In this paper we consider the setting where the regressor is a functional data such as a curve or an image and the response is a scalar. We propose the “functional smooth lasso” (FSL) approach to simultaneously regularize the roughness and the size of the nonzero regions of the functional linear regression estimates. An efficient algorithm is developed for computing FSL. The degrees of freedom of FSL is derived and incorporated into the automatic tuning of regularization parameters. Furthermore, we prove the consistency and the convergence rate of FSL. An interesting finding is that the convergence rate depends on the degree of the “smoothness” of the predictors. The proposed method is illustrated via simulation studies and real data application.

# **Multi-Dimensional Trimming Based on Data Depth**

Yijun Zuo

Department of Statistics and Probability, Michigan State University, East Lansing

With a natural order principle, trimming in one dimension is straightforward. One-dimensional trimmed means are among the most popular estimators of the center of data and have been used in various fields of statistics and in our daily life. Trimmed means can overcome the high sensitivity of the mean to outliers (or heavy-tailed data) and the low efficiency of the median for light-tailed data. Hence they can serve as compromises between the mean and the median.

Multi-dimensional data often contain outliers, which typically are far more difficult to detect than in one dimension. A robust procedure such as the multi-dimensional trimming that can automatically detecting outliers or “heavy tails” is thus desirable. The task of trimming, however, becomes non-trivial, for there is no natural order principle in high dimensions. In this talk, multi-dimensional trimming based on “data depth” is discussed. It is found that depth-trimmed means can possess very desirable properties such as high efficiency and high robustness. Furthermore, inference procedures based on the depth-trimmed means can outperform the classical Hotelling’s  $T^2$  (and the univariate  $t$ ) ones. Applications of data depth trimming such as clustering and dimension reduction are also addressed. Contributions of Professor Bickel to trimming are discussed.