Mathematisches Forschungsinstitut Oberwolfach

Report No. 39/2009

DOI: 10.4171/OWR/2009/39

Challenges in Statistical Theory: Complex Data Structures and Algorithmic Optimization

Organised by
Rudolf J. Beran (Davis, CA)
Claudia Klüppelberg (TU München)
Wolfgang Polonik (Davis, CA)

August 23rd – August 29th, 2009

Abstract. Technological developments have created a constant incoming stream of complex new data structures that need analysis. Modern statistics therefore means mathematically sophisticated new statistical theory that generates or supports innovative data-analytic methodologies for complex data structures. Inherent in many of these methodologies are challenging numerical optimization methods. The proposed workshop intends to bring together experts from mathematical statistics as well as statisticians involved in serious modern applications and computing. The primary goal of this meeting was to advance the mathematical and methodological underpinnings of modern statistics for complex data. Particular focus was given to the advancement of theory and methods under non-stationarity and complex dependence structures including (multivariate) financial time series, scientific data analysis in neurosciences and bio-physics, estimation under shape constraints, and high-dimensional discrimination/classification.

Mathematics Subject Classification (2000): primary: 62xx, *secondary:* 60Bxx, 60Gxx, 60Hxx, 60Jxx, 90Cxx, 92xx.

Introduction by the Organisers

The workshop *Challenges in statistical theory: Complex data structures and algorithmic optimization*, organised by Rudolf Beran (Davis, CA), Claudia Klüppelberg (TU München) and Wolfgang Polonik (Davis, CA) was held August 23rd – August 29th, 2009. This meeting was well attended by 49 participants with diverse geographic, demographic and disciplinary representation.

The theme of the conference addressed the challenges to modern Statistics created by the ongoing emergence of novel, large, and complex data types. Human ability to collect data through sophisticated electronic technologies has outstripped human ability to distill information from the data. Resolving the situation requires, among other things, fundamental new developments in statistical theory and algorithms. Traditional probabilistic studies of statistical procedures remain an important tool but no longer suffice. Data is arguably not random in the sense of probability theory; data may reside naturally on a manifold or other algebraic structure; the procedure under study may be very complicated; and probability modeling in some modern problems, such as classification of highly structured data, has not been effective. Emerging new types of data include: single molecule observations; complex simultaneous recording of several neurons; the outcomes of computer experiments; high dimensional observations of brain waves that need to be processed in real time (if possible); or high-frequency financial data.

To date, the most powerful statistical methodologies have been developed for data that resides in a Euclidean space. Emerging data types pose a variety of questions that include: On what algebraic structure does the data naturally reside? On this algebraic structure, can we develop statistical methodologies that address the questions posed by those who collected the data? In particular, can we devise for such data analogs of successful statistical methodologies for Euclidean data? Meeting such challenges requires communication among those most involved with the new types of data, those with the expertise to identify suitable mathematical formulations, those who have thought deeply about abstract statistical inference, and those who seek to devise new paradigms for statistical reasoning beyond probability modeling of the data.

Thus, a fundamental task for Statistics is to develop powerful theoretical tools that engender and validate effective methodologies for the analysis of modern data types arising in a variety of fields. To do so first requires gaining sufficient disciplinary and mathematical insight into the new underlying data structures. The workshop brought together experts from mathematical statistics and the statistical sciences. The primary goal was to address the foregoing challenges by broadening the mathematical underpinnings of modern statistics. The secondary goal was to foster cross-fertilization between the core of statistics and the statistical sciences.

Workshop participants presented a range of novel data structures and of methodologies proposed for their analysis. Mathematical advances were exhibited, for instance in the area of model selection for functions in high dimensional spaces, in estimation under heavy tails, or in estimation under shape restrictions. Tools, such as Malliavin calculus, for the large sample analysis of certain statistics were discussed. Methodological advances in the areas of modeling of nonstationarity, the construction of confidence intervals for classification error, or the testing of functional autoregression were treated. Algorithmic and/or computational issues are inherent in many of these challenges, and many of the presentations addressed this aspect.

Summary: Complex high dimensional data is rather the norm than the exception in modern statistics, and modeling or analyzing such complex data is a huge challenge. In order to properly understand these approaches effective mathematical techniques are necessary. Making advanced methodology feasible in practical applications usually also requires devising sophisticated optimization/algorithmic methodologies. New paradigms beyond probability modeling are needed to validate complicated statistical procedures. Balancing all of these ingredients is a fundamental challenge. Substantial progress in these directions requires input from various sides. The workshop brought out the issues and made significant contributions to the program outlined.

Workshop: Challenges in Statistical Theory: Complex Data Structures and Algorithmic Optimization

Table of Contents

Abstracts

Modeling Nonstationarity using Deformed Random Fields

Ethan B. Anderes

This talk will present work on using non-linear spatial transformations to model certain types of nonstationarity. The use of deformations for modeling nonstationary processes has been applied in diverse fields from geophysics to image analysis. These models are natural extensions of stationary processes that are simple to understand but give rise to a diverse range of behavior. Even though these models seem a good choice when modeling nonstationary random fields they are generally difficult to work with because of the complex restrictions on the deformations like invertablility. We will begin the talk by discussing general approaches to modeling nonstationarity, then focus attention to the deformation approach. Finally we review some open problems for classifying nonstationary random fields.

The use of deformations to model nonstationary processes was first introduced to the spatial statistics literature by Sampson and Guttorp [5]. Their work, as well as that of subsequent authors consider estimating the deformation f when observing a deformed random field $Z \circ f$ at sparse observation locations with independent replicates of the random field. Three recent papers ([4], [2],[3]) study the different problem of estimating a deformation f from dense observations of a single realization of a deformed isotropic random field $Z \circ f$ in two dimensions. These deformed isotropic random fields provide a flexible semi-parametric model of nonstationarity for random fields. In addition, this observation scenario is becoming increasingly important with the abundance of high resolution digital imagery and remote sensing. One of the more recent motivations for the deformation model under the one-realization observation scenario is gravitational lensing of the cosmic microwave background (CMB). The gravitational effect from intervening matter distort the CMB images to produce deformed random eld observations. In the hope of improving estimates of cosmological parameters and the mass distribution in the universe (including dark matter) there is considerable interest in detecting and measuring the lensing of the CMB

We present recent work on establishing the strong consistency for the estimation of the deformation f when observing $Z \circ f$ on a dense grid in a bounded simply connected domain in $\mathbb{R}^2$, as the grid spacing goes to zero [2]. We also present fixed domain asymptotic results that establish consistent estimates of the variance and scale parameters for a Gaussian random field with a geometric anisotropic Matérn autocovariance in dimension $d > 4$ (see [1]). When $d < 4$ this is impossible due to the mutual absolute continuity of Matérn Gaussian random fields with different scale and variance (see Zhang [6]). Informally, when $d > 4$, we show that one can estimate the coefficient on the principle irregular term accurately enough to get a consistent estimate of the coefficient on the second irregular term. These two coefficients can then be used to separate the scale and variance. We extend our results to the general problem of estimating a variance and geometric

anisotropy for more general autocovariance functions. Our results illustrate the interaction between the accuracy of estimation, the smoothness of the random field, the dimension of the observation space, and the number of increments used for estimation. As a corollary, our results establish the orthogonality of Matérn Gaussian random fields with different parameters when $d > 4$. The case $d = 4$ is still open.

References

[1] E. Anderes , *On the consistent separation of scale and variance for Gaussian random fields*, The Annals of Statistics, *to appear*.

[2] E. Anderes and S. Chatterjee, *Consistent estimates of deformed isotropic Gaussian random fields on the place*, The Annals of Statistics **37** (2009), 2324–2350.

[3] E. Anderes and M. Stein, *Estimating Deformation of Isotropic Gaussian Random Fields on the Plane*, The Annals of Statistics **36** (2008), 719–741.

[4] M. Clerc, and S. Mallat, *Estimating deformations of stationary processes*, The Annals of Statistics **31** (2003), 1772–1821.

[5] P. Sampson and P. Guttorp. *Nonparametric estimation of nonstationary spatial covari- ance structure*, Journal of the Americal Statistical Association **87** (1992), 108-119.

[6] H. Zhang *Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics*, Journal of the Americal Statistical Association **99** (2004), 250-261.

Model selection for functions on high-dimensional spaces

Lucien Birgé

(joint work with Yannick Baraud)

We observe some random object X with distribution $P(s, \tau)$ which depends on an unknown parameter $s \in \mathcal{S}$ and a known information parameter $\tau > 0$ which converges to zero when the amount of information goes to infinity. For instance $\tau = n^{-1}$ when X is an i.i.d. sample of size n, and $\tau = \sigma^2$ when X is a white noise with variance σ^2. For simplicity, we assume here that $\tau < 1/2$. We also assume that $\mathcal{S}$ is a subset of $\mathbb{L}_2(\mu)$ with μ a probability on $E = [-1, 1]^k$ and we measure the risk $\mathcal{R}(s, \hat{s}, \tau)$ of an estimator $\hat{s}(X)$ at s using the quadratic loss, i.e.

$$\mathcal{R}(s, \hat{s}, \tau) = \mathbb{E}\left[\|\hat{s}(X) - s\|^2\right].$$

When $k = 1$ and we assume that s is smooth, for instance $s \in \mathcal{H}_1(\beta)$, the space of Hölderian functions on $[-1, 1]$ with smoothness β, we can build estimators with a risk bounded by $C(s, \beta)\tau^{2\beta/(2\beta+1)}$. In this case we get a reasonable rate (for instance $\tau^{2/3}$ for Lipschitz functions) unless β is small. But the situation changes drastically when k gets large, the assumption that $s \in \mathcal{H}_k(\beta)$ leading to the conclusion that one can only get a risk bounded by $C(s, \beta)\tau^{2\beta/(2\beta+k)}$. This is a very weak bound when k is large, unless β is also very large, which is often unrealistic. Smoothness assumptions are useful in low dimensions but not any more in large dimensions (a phenomenon which is often refered to as the "curse of dimensionality"). They have therefore to be replaced by more useful ones. A simple choice is to assume a parametric model for s but since this is quite a strong

assumption, more flexible structural assumptions on s have been considered in the past like additive models, the single index model, artificial neural networks, etc. Many such structural assumptions amount to write s as $g \circ u$ where u is a bounded function on E with a rather simple structure and g is a function on the line. In this case, the simple structure of u allows to estimate it at a high rate and the estimation difficulty is concentrated on g which is a function on a one-dimensional space. Thus we avoid to mix a complicated structure with a high dimension.

Recently, some attempts have been made to compute risk bounds for the estimation of parameters of the form $g \circ u$. The first one is [4] (already announced by O. Lepski in a conference at the CIRM and available to the authors several years ago) but it only deals with the white-noise framework and the $\mathbb{L}_\infty$-loss which happens to lead to quite different results. The second one [3] is closer to our approach but focuses on some special cases of composite functions and the regression framework. Our approach, based on model selection has actually more general purposes:

- It applies to various statistical models including density estimation, estimation of a regression function or the intensity of a Poisson process, estimation for Gaussian sequences, etc.
- It is completely adaptive to various smoothness or structural assumptions.
- It is robust in the sense that we only assume that $g \circ u$ is an approximation for s.

Unfortunately, such a level of generality has a price and our construction leads to very complex estimators which are unlikely to be computed in a reasonable amount of time.

The main ingredient for our construction is the following theorem that applies to various statistical frameworks as indicated above and proved in [1] and [2].

Theorem 1. *Let $\mathbb{S}$ be a countable family of subsets S of $\mathbb{L}_2(\mu)$, each S being a subset of some finite-dimensional linear space with dimension $\mathcal{D}(S) \geq 1$, and let π be some probability on $\mathbb{S}$. There exists an estimator $\hat{s} = \hat{s}(X)$ with values in $\bigcup_{S \in \mathbb{S}} S$ satisfying, for all $s \in \mathcal{S}$,*

$$\mathbb{E}\left[\|\hat{s} - s\|^2\right] \leq C \inf_{S \in \mathbb{S}} \left\{ \inf_{t \in S} \|t - s\|^2 + \tau \left[\mathcal{D}(S) - \log\left(\pi(S)\right)\right] \right\}, \tag{1}$$

for some universal constant C.

Starting from this theorem and assuming that g is continuous, we introduce families of subsets of finite-dimensional linear spaces, $\mathbb{F}$ and $\mathbb{T}$, to approximate g in $\mathbb{L}_\infty(\mu)$-norm and u in $\mathbb{L}_2(\mu)$-norm respectively. We use these two families to build a new family $\mathbb{S}$ of models which approximate $g \circ u$ and we apply Theorem 1 to $\mathbb{S}$ to get an estimator $\hat{s}$ for s with a risk bound of the form (1). Apart from an unavoidable bias term of the order of $\|s - g \circ u\|^2$, the bound involves the properties of the families of models $\mathbb{F}$ and $\mathbb{T}$ and the modulus of continuity of g. To be more specific, when $s = g \circ u$ and g is Lipschitz, the resulting risk bound is the same (up to an extra $\log(\tau^{-1})$ factor) as the sum of the risks for estimating g with the family $\mathbb{F}$ and u with the family $\mathbb{T}$ by the method of Theorem 1. One possible illustration is the estimation of $g \circ u$ when u takes its values in $[-1, 1]^l$, $l < k$, and

both u and g have isotropic Hölderian smoothnesses of order α and β respectively, with $\alpha > \max\{\beta; 1\}$. Then, the overall smoothness of $s = g \circ u$ is β which results in a risk bound of order $\tau^{2\beta/(2\beta+k)}$ when we consider classical methods based on smoothness assumptions to estimate s. Our strategy which uses the fact that s is a composite function results in a risk bound of order

$$\max\left\{\tau^{2\beta/(2\beta+l)}; \left[\tau \log(\tau^{-1})\right]^{2\theta/(2\theta+k)}\right\} \quad \text{with } \theta = \alpha(\min\{\beta; 1\}).$$

It is easy to see that, under our assumption on α, this bound provides a better rate of convergence to zero than $\tau^{2\beta/(2\beta+k)}$ when $\tau \to 0$. Moreover we need not know in advance the values of l, α and β to build our estimator.

The method actually also applies to approximations of s by additive (or generalized additive) models, the single or multiple index model, artificial neural networks and we may actually design an estimator which automatically finds the best of these structures to approximate s.

References

[1] L. Birgé, *Model selection via testing: an alternative to (penalized) maximum likelihood estimators*, Ann. Inst. Henri Poincaré, Probab. et Statist., **42** (2006), 273–325.

[2] L. Birgé, *Model selection for Poisson processes*, in Asymptotics: particles, processes and inverse problems, Festschrift for Piet Groeneboom (E. Cator, G. Jongbloed, C. Kraaikamp, R. Lopuhaä and J. Wellner, eds), IMS Lecture Notes – Monograph Series **55** (2007), 32–64.

[3] J. Horowitz and E. Mammen, *Rate-optimal estimation for a general class of nonparametric regression models with unknown link function*, Ann. Statist. **35** (2007), 2589–2619.

[4] A.B. Juditsky, O.V. Lepski and A.B. Tsybakov, *Nonparametric estimation of composite functions*, Ann. Statist. **37** (2009), 1360–1404.

Estimation of high-dimensional intervention and causal effects

Peter Bühlmann

We assume that we have observational data, generated from an unknown underlying directed acyclic graph (DAG) model. A DAG is typically not identifiable from observational data, but it is possible to consistently estimate the equivalence class of a DAG. Moreover, for any given DAG, causal effects can be estimated using intervention calculus. Here, we combine these two parts. For each DAG in the estimated equivalence class, we use intervention calculus to estimate the causal effects of the covariates on the response. This yields a collection of estimated causal effects for each covariate. We show that the distinct values in this set can be consistently estimated by a new algorithm that uses only local information of the graph. Sparsity and so-called faithfulness for the distribution are the two key assumptions for the asymptotic analysis which also covers the framework with many more variables than sample size. Our local approach is computationally fast and feasible in high-dimensional problems.We demonstrate the merits of our methods on two large-scale biological systems.

Our work is motivated by the following problem in biology. We want to know which genes play a role in a certain phenotype, say a disease status or, in one of our

cases, a continuous value of riboflavin (vitamin B2) production in the bacterium Bacillus Subtilis. To be more precise, our goal is to infer which genes have an effect on the phenotype in terms of an intervention: if we knocked down single genes, which of them would show a relevant or important effect on the phenotype? The difficulty is, however, that the available data are only observational. Using such observational data, we want to infer all (single gene) intervention effects. This task coincides with inferring causal effects, a well-established area in statistics, cf. [3] or [4]. We emphasize that in our applications, it is exactly the intervention or causal effect which is of interest, rather than a regression-type effect of association.

[3, p.285] formulates the distinction between associational and causal concepts as follows: An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. (...) Every claim invoking causal concepts must be traced to some premises that invoke such concepts; it cannot be inferred or derived from statistical associations alone. Thus, in order to obtain causal statements from observational data, one needs to make additional assumptions. One possibility is to assume that the data were generated by a directed acyclic graph (DAG) which is known beforehand. DAGs describe causal concepts, since they code potential causal relationships between variables: the existence of a directed edge $x \to y$ means that x may have a direct causal effect on y, and the absence of a directed edge $x \to y$ means that x cannot have a direct causal effect on y.

Given a set of conditional dependencies from observational data and a corresponding DAG model, one can compute causal effects using intervention calculus ([3]).

Here, we consider the problem of inferring causal information from observational data, under the assumption that the data were generated by an unknown DAG. This is a more realistic assumption, since in many practical problems, one does not know the DAG. In this scenario, the causal effect is typically not defined uniquely, and that is not surprising given the description of causality by [3] above.

A DAG is typically not identifiable from observational data, because conditional dependencies only determine the skeleton and the so-called v-structures of the graph. The skeleton and v-structures determine an equivalence class of DAGs that all correspond to the same probability distribution. This equivalence class, which is identifiable from observational data, can be described by a completed partially directed acyclic graph (CPDAG).

We describe a new, computationally feasible algorithm, even if the number of variables (i.e. nodes in the graph) is large, which uses the CPDAG as input for inferring lower bounds on intervention or causal effects. Furthermore, we show that in the case of noise and estimation error, we can still asymptotically infer the CPDAG and the lower bounds for causal effects even if the number of variables p (number of nodes in the graph) is much larger than sample size n, $p \gg n$. Such a consistency result relies on sparsity of the (causal) DAG and the so-called faithfulness assumption for the data-generating probability distribution with respect

to the underlying DAG. Details are given in [2] and some of the results there rely on [1]. Furthermore, we demonstrate the method to predict the most important intervention effects in two large-scale biological systems from Bacillus Subtilis and S.Cerevisiae.

References

[1] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, Journal of Machine Learning Research **8** (2007), 613–636.
[2] M.H. Maathuis, M. Kalisch, and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data*, Annals of Statistics **37** (2009), 3133–3164.
[3] J. Pearl, *Causality: models, reasoning and inference*, Cambridge University Press, 2000.
[4] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, The MIT Press, 2nd edition, 2000.

Distilled Sensing: Active Sensing for Sparse Signal Recovery

Rui M. Castro
(joint work with Jarvis D. Haupt, Robert D. Nowak)

The study and use of sparse representations in data-rich applications has garnered significant attention in the signal processing, statistics, and machine learning communities. In the present work we develop a novel sensing procedure called *Distilled Sensing* (DS), which is a sequential and adaptive approach for recovering sparse signals in noise.

Passive sensing approaches, currently the most widespread data collection methods, involve non-adaptive data collection procedures that are completely specified before any data is observed. In contrast, DS collects data in a sequential and adaptive manner. Often such procedures are known as *active sensing* or *sequential experimental design*, and allow the use of data observed in earlier stages to guide the collection of future data. The added flexibility of active sensing, together with a sparsity assumption, has the potential to enable extremely efficient and accurate inference.

In essence, DS exploits the fact that it is often easier to rule out locations where the signal is absent than it is to precisely detect the location of non-zero signal components. Following each observation of the sequential DS procedure, a coarse-level significance test is performed to effectively identify a lower-dimensional subspace containing the unknown signal. This allows the sensing procedure to carefully focus on the relevant signal subspace, gradually *distilling* the observations being made, and resulting in rather dramatic improvements in the recoverability of sparse signals compared to that of passive sensing methods. Similar sensing strategies have been previously proposed in the bio-statistics literature, and the work of Zehetmayer *et al* is representative of those approaches. The main contribution of our work is the formal quantification of the gains attained by such procedures.

Specifically, we consider the following observation model, which is suitable for a variety of applications including the monitoring of the radio spectrum for opportunistic transmission and astronomical surveying. Let $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$

be a sparse vector, such that most entries are zero. We cannot observe $\boldsymbol{x}$ directly, but instead make observations of the form

$$Y_i^{(j)} = x_i + \left(\gamma_i^{(j)}\right)^{-1/2} Z_i^{(j)} \ , j = 1, \ldots, k$$

where $Z_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are Gaussian random variables with mean zero and variance one, and $\gamma_i^{(j)}$ is the precision associated with the jth measurement of the ith entry of $\boldsymbol{x}$ (by convention if $\gamma_i^{(j)} = 0$ the entry is not measured). In addition we impose a restriction on the *total precision*, namely $\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_i^{(j)} \leq n$.

In the work of Donoho and Jin the authors considered a very particular instance of this model, where $k = 1$ and therefore only a single measurement step $\boldsymbol{X}^{(1)}$ is taken, with $\gamma_i^{(1)} = 1$ for all i. They show the following result in that setting:

Theorem [Donoho and Jin 2004]: *Let $k = 1$ and $\boldsymbol{x} \geq 0$ be a sparse vector such that only $n^{1-\beta}$ of the entries are non-zero ($\beta \in (0,1)$). Let $S = \{i : x_i \neq 0\}$ denote the support of the signal. If the non-zero entries of $\boldsymbol{x}$ are larger than $\sqrt{2\beta \log n}$ then there is a support set estimator $\widehat{S}$ obtained by thresholding the observations $\boldsymbol{Y}^{(1)}$ such that*

$$\mathrm{FDP}(\widehat{S}) \xrightarrow{P} 0 \text{ and } \mathrm{NDP}(\widehat{S}) \xrightarrow{P} 0, \quad \text{as } n \to \infty \ ,$$

where $\mathrm{FDR}(\widehat{S}) = |\widehat{S} \setminus S|/|\widehat{S}|$ *is the* False Discovery Proportion *and* $\mathrm{NDR}(\widehat{S}) = |S \setminus \widehat{S}|/|S|$ *is the* Non Discovery Proportion. *Conversely if the non-zero entries of $\boldsymbol{x}$ are smaller than $\sqrt{2\beta \log n}$ no procedure can simultaneously control the FDP and NDP.*

This result shows that the signal support can only be reliably recovered if the signal magnitude is larger than $\sqrt{2\beta \log n}$. Contrasting with this, we show that if multiple measurement steps are allowed (under a total precision budget) a much better result is achievable. This improvement is only possible when $\boldsymbol{\gamma}^{(j)}$ are allowed to depend explicitly on past observations $\{\boldsymbol{Y}^{(\ell)}, \boldsymbol{\gamma}^{(\ell)}\}_{\ell<j}$. If such dependence in not allowed there is no advantage of taking multiple measurements, and the earlier result the best possible. For the proposed measurement model we show that:

Theorem: [Haupt, Castro and Nowak 2009]: *Let $\boldsymbol{x} \geq 0$ with $n^{1-\beta}$ non-zero components of amplitude $\mu(n)$ ($\beta \in (0,1)$). There exists a sequential measurement procedure (called Distilled Sensing) using $k = 1 + \log\log n$ measurement steps and satisfying the precision budget $\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_i^{(j)} \leq n$, yielding a support estimate $\widehat{S}_{DS}$ such that if $\mu(n) > \log \ldots \log n$ for some finite iteration of the logarithm then*

$$FDP(\widehat{S}_{DS}) \xrightarrow{P} 0, \quad NDP(\widehat{S}_{DS}) \xrightarrow{P} 0, \qquad \text{as } n \to \infty \ .$$

Furthermore the performance gains above are achievable only when allowing adaptation of the measurement procedure based on previous observations.

This result shows that by using a sequential experimental design approach it is possible to greatly enlarge the class of signals that can be reliably recovered. Similar results can also be stated for detection of sparse signals, in the spirit of the work of Ingster (2007). A preliminary version of this work has appeared in Haupt *et al* 2009.

References

[1] D. Donoho and J. Jin, *Higher Criticism for Detecting Sparse Heterogeneous Mixtures*, Ann. Statist., **32**(3), 962–994.

[2] S. Zehetmayer, P. Bauer and M. Posch, *Optimized multi-stage designs controlling the false discovery or the family-wise error rate*, Statist. Med. 2008; **27**, 4145–4160

[3] J. Haupt, R. Castro, and R. Nowak, *Distilled sensing: selective sampling for sparse signal recovery.* in 12th Conference on Articial Intelligence and Statistics, Clearwater Beach, Florida, April 2009.

Application of Heteroskedastic Spatial Models to Computer Experiments

Richard A. Davis

(joint work with Jay Breidt, Wenying Huang, Ke Wang)

We consider modeling a deterministic computer response as a realization from a stochastic heteroskedastic process (SHP), which incorporates a spatially-correlated volatility process into the traditional spatially-correlated Gaussian process (GP) model. Unconditionally, the SHP is a stationary non-Gaussian process, with stationary GP as a special case. Conditional on a latent process, the SHP is a non-stationary GP. The sample paths of this process offer more modeling flexibility than those produced by a traditional GP, and can better reflect prediction uncertainty. GP prediction error variances depend only on the locations of inputs, while SHP can reflect local inhomogeneities in a response surface through prediction error variances that depend on both input locations and output responses.

We use maximum likelihood for inference, which is complicated by the high dimensionality of the latent process. Accordingly, we develop an importance sampling method for likelihood computation and use a low-rank kriging approximation to reconstruct the latent process. Responses at unobserved locations can be predicted using empirical best predictors or by empirical best linear unbiased predictors. Prediction error variances are also obtained. In examples with simulated and real computer experiment data, the SHP model is superior to traditional GP models.

On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models

Sonja Greven
(joint work with Thomas Kneib)

Linear mixed models are a powerful inferential tool in modern statistics and have been used in a wide range of areas. Using penalized splines, they are employed in nonparametric regression to model smooth functions, varying coefficients or surfaces. In functional data analysis, they have been used in functional principal component analysis. They allow for simple additive extension of such models as well as the extension to multilevel or longitudinal data. Moreover, they are computationally attractive for large and complex data sets. We consider the linear mixed model

$$y = X\beta + Zb + \varepsilon,$$

where X and Z are known design matrices, β is a fixed parameter vector, b and ε are assumed to be independent, $b \sim N(0,D)$ and $\varepsilon \sim N(0,\sigma^2 I_n)$.

As linear mixed models offer large flexibility in modeling, model selection becomes increasingly important. Selection here includes the selection of random effects, such as those modeling nonlinearity of a smooth function in the penalized splines approach. In standard settings, the Akaike information criterion (AIC) is defined as minus twice the maximized log-likelihood, plus $2k$, two times the number of estimable parameters in the model. The AIC is asymptotically unbiased for twice the expected relative Kullback-Leibler distance, and minimizing the AIC thus can be seen as minimizing the average distance between an approximating model and the underlying truth. In the linear mixed model, two versions of the AIC have been used for model selection.

The marginal AIC (mAIC) is based on the marginal likelihood derived from $y \sim \mathcal{N}(X\beta, \sigma^2 I_n + ZDZ')$, and uses the number of fixed effects plus the number of variance components as the number of parameters. Note that this is the AIC returned by standard software such as R and SAS. We show that the marginal AIC is not asymptotically unbiased for twice the expected relative Kullback-Leibler distance. It favors smaller models without random effects. This is due to standard regularity conditions not being fulfilled in mixed models. First, the parameter space is not (a transformation of) $\mathcal{R}^k$, as the space for the variance components is restricted to ensure that D is positive semi-definite. Second, observations are not independent, as the random effects induce a correlation structure in y.

The conditional AIC (cAIC) was introduced by Vaida and Blanchard [3] as more appropriate when the focus is on the random rather than on the fixed effects. The cAIC uses the conditional likelihood based on $y|b \sim \mathcal{N}(X\beta + Zb, \sigma^2 I_n)$, as well as the effective degrees of freedom, i.e the trace of the hat matrix. Under the assumption that $\sigma^{-2}D = D_*$ is known, they show that the cAIC is unbiased for the conditional Akaike information, a conditional version of twice the expected relative Kullback-Leibler distance. They propose to use the cAIC with estimated D_* when it is not known. We show that ignoring the estimation uncertainty in

$\widehat{D}_*$ leads to the following peculiar behavior: the cAIC always chooses inclusion of an additional random effect, as long as the random effect is not predicted to be exactly zero.

A corrected version of the conditional AIC taking into account the estimation uncertainty has been proposed by Liang et al [2]. However, this corrected cAIC had not been available in closed form so far. Numerical approximations were intractable for large data sets. We now derive a closed form representation of the corrected cAIC, which can be computed efficiently.

Theorem 1 For the cAIC in the linear mixed model with unknown $D_*(\theta_*)$, the effective degrees of freedom Φ_0 can be written as

$$\Phi_0 = \text{trace}\left(\frac{\partial \widehat{y}}{\partial y}\right) = n - \text{trace}(\widehat{A}_*) + \sum_{j=1}^{s} e_j' \widehat{B}_*^{-1} \widehat{G}_* \widehat{A}_* \widehat{W}_{*,j} \widehat{A}_* y,$$

assuming that (after reordering) $\theta_* = (\theta_s', \theta_t', \theta_{q-s-t}')'$ for some $0 \leq s, t \leq q$, such that $\Theta = \{\theta_* | \theta_s \in \Theta_s \subseteq \mathcal{R}^s, \theta_t \in C \subset \mathcal{R}^t, \theta_{q-s-t} \in F(\theta_s, \theta_t) \subset \mathcal{R}^{q-s-t}\}$, $\widehat{\theta}_s$ is in the interior of Θ_s, C is a cone with vertex at some θ_0, $F(\theta_s, \theta_0) = \theta_1$ for all θ_s for some θ_1, and $(\widehat{\theta}_t, \widehat{\theta}_{q-s-t}) = (\theta_0, \theta_1)$.

$$A_* = V_*^{-1} - V_*^{-1} X (X' V_*^{-1} X)^{-1} X' V_*^{-1},$$
$$W_{*,j} = \frac{\partial}{\partial \theta_{*,j}} V_*, \quad U_{*,jl} = \frac{\partial}{\partial \theta_{*,l}} W_{*,j}, \quad j, l = 1, \ldots, s,$$

e_j is the jth unit vector, $V_* = \sigma^{-2}\text{Cov}(y)$, $\widehat{B}_*$ is negative definite, and the entries in G_* and B_* involve simple expressions in y, A_*, $W_{*,j}$ and $U_{*,jl}$, $j, l = 1, \ldots, s$. Hats denote matrices as functions of θ_* evaluated at the (restricted) maximum likelihood estimator $\widehat{\theta}_*$.

We illustrate all results in simulation studies and in an application of additive mixed models to the analysis of childhood malnutrition in Zambia.

A longer technical report is available at [1].

References

[1] S. Greven and T. Kneib, *On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models*, Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 179 (2009). http://www.bepress.com/jhubiostat/paper179

[2] H. Liang, H. Wu and G. Zou, *A note on conditional AIC for linear mixed-effects models*, Biometrika **95** (2008), 773–778.

[3] F. Vaida and S. Blanchard, *Conditional Akaike information for mixed-effects models*, Biometrika **92** (2005), 351–370.

A fast and memory-efficient boosting implementation for generalized linear and additive models

Torsten Hothorn

Boosting can be seen as a very general functional approach to statistical model fitting. Its flexibility is extremely attractive also from a computational point of view, since a huge class of classical and modern statistical models can be fitted by such a procedure.

Two model classes are especially interesting; generalized linear models and generalized additive models. [3] introduced boosting as a means for fitting additive models. For each model component, one base-learner is specified and only one base-learner is selected and updated in each iteration of the algorithm. For linear models, the same idea can be applied [1]. [2] give an overview on these and related issues. The generality of the additive modelling framework lead to boosting algorithms for structured additive models, especially useful for modelling space-time data and geoadditive regression [5].

The conceptual tools for boosting well-defined statistical models are now in place. However, there is always a trade-off between computational flexibility, generality and efficiency of a specific implementation of such models. Here, we focus on generalized linear and additive models for problems where both the number of observations and exploratory variables may be in the millions. We present techniques to speed up computations and to reduce the memory footprint considerably.

For smooth model components, P-splines (univariate) or tensor-product P-splines (interaction surfaces) are used as base-learners. This choice is computationally attractive because the B-spline design matrices are sparse by definition. Thus, fitting such base-learners can make use of sparse matrix functionality. Moreover, it is possible to remove tied observations before iteratively fitting these base-learners. Thus, fitting a smooth base-learner only requires fitting a rather low-dimensional system of equations.

Experiments suggest that high-dimensional linear models can be fitted by a component-wise boosting algorithm really fast (even faster than the lasso or elastic-net). Additive models for millions of observations can be fitted on standard desktop computers. Bootstrapping for model tuning and model inference benefits from these improvements as well. The corresponding R package **mboost** [4] is available from `R-forge.R-project.org`.

References

[1] Peter Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34 (2):559–583, 2006.

[2] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[3] Peter Bühlmann and Bin Yu. Boosting with L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–338, 2003.

[4] Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. *mboost: Model-Based Boosting*, 2009. URL `http://R-forge.R-project.org`. R package version 2.0-0.

[5] Thomas Kneib, Torsten Hothorn, and Gerhard Tutz. Variable selection and model choice in geoadditive regression models. *Biometrics*, 65(2):626–634, 2009.

Asymptotically Optimal Spatially Adaptive Splines

Xiaoming Huo
(joint work with Heeyoung Kim)

In penalized splines [1], generalized cross validation (GCV) is widely used to choose the algorithmic parameter. It is known that GCV is asymptotically optimal (AO): under certain conditions, when the sample size goes to infinity, the estimates that corresponds to GCV is as good as the best estimates; Because its inefficiency converges to the best possible (which is one) [2]. Classical GCV consider a univariate penalizing parameter; while spatially adaptive GCV (SA-GCV) considers varying penalty parameters over the space [3, 4]. We give a sufficient condition on the AO of SA-GCV. Our work automatically leads to some design guidelines. We name the resulting splines the *asymptotically optimal spatially adaptive splines.* The key mathematical ingredient in this work is to work out conditions, under which the coefficient of variations of the reciprocals of the eigenvalues of a design matrix goes to infinite when the sample size goes to infinity.

References

[1] Green, P. J. and Silverman, B. W. (1994), *Nonparametric regression and generalized linear models : a roughness penalty approach,* Chapman & Hall, New York.
[2] Li, K.-C. (1986), *Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing,* The Annals of Statistics, **14**, 1101-1112.
[3] Pintore, A., Speckman, P., and Holmes, C. C. (2006), *Spatially adaptive smoothing splines*, Biometrika, **93**, 113-125.
[4] Ruppert, D. and Carroll, R. J. (2000), *Spatially-adaptive penalties for spline fitting*, Australian and New Zealand Journal of Statistics, **42**, 205-223.

Testing the stability of the functional autoregressive process

Marie Hušková
(joint work with Lájos Horváth, Piotr Kokoszka)

The talk concerns test procedures for detection of a changes in a functional autoregressive process. We consider the functional autoregressive process defined by the equation

$$X_{n+1} = \Psi_n X_n + \varepsilon_{n+1},\ n = 1, \ldots, N,$$

where $X_n \in L_2([0,1])$ are observed curves, ε_n are independent identically distributed (iid) mean zero innovations in $L_2([0,1])$ and Ψ_n are operators. We propose a method for testing the constancy of the operators Ψ_n:

$$H_0 : \Psi_1 = \ldots = \Psi_N$$

against a change point alternative

$$H_1 : there \quad is \quad k^* < N : \quad \Psi_1 = \ldots = \Psi_{k^*} \neq \Psi_{k^*+1} = \ldots = \Psi_N.$$

The developed test procedures are based on the differences of the sample autocovariances of projections of the functional observations on estimated principal components(PCs). The limit distribution can be derived by replacing the estimated PCs by their population counterparts and using a functional central limit theorem for ergodic sequences. But in the functional setting, this replacement introduces asymptotically nonnegligible terms, which cancel because of the special form of the test statistic. The estimated PCs are determined only up to a sign, and our statistic is invariant to these random signs. Finally, to show that the remaining terms due to the estimation of the PCs are asymptotically negligible, we develop a new technique which involves the truncation at lag $O(logN)$ of the moving average representation of the ARH(1) process , a blocking technique that utilizes this truncation (and Mensovs inequality). We think that these tools will prove useful in other inference problems related to the functional ARH(1) model. Limit behavior of the developed test statistic is investigated.
Finite sample performance is examined by an application to a data set.

References

[1] L. Horváth, M. Hušková and P. Kokoszka, *Testing the stability of the functional autoregressive process*, to appear in the Journal of Multivariate Analysis (2009).

On Riesz and Wishart Distributions Associated with Decomposable Undirected Graphs

Thomas Klein
(joint work with Steen A. Andersson)

Classical Wishart distributions on the open convex cones of positive definite matrices and their fundamental features are extended to generalized Riesz and Wishart distributions associated with decomposable undirected graphs using the basic theory of exponential families. The families of these distributions are parameterized by their expectations/natural parameter and multivariate shape parameter and have a non-trivial overlap with the generalized Wishart distributions defined in [1, 2]. This work also gives an explicit description of the "Hyper Wishart laws" introduced in [3] and extends the "Wishart distributions of type I" from [4]. We shall emphasize that our main motivation for defining and investigating generalized Riesz/Wishart distributions lies in the fact that they form a natural, flexible, and tractable distribution family adapted to the structure of the underlying sample space and well-suited for likelihood inference. Moreover, we present various examples of how generalized Riesz/Wishart distributions appear naturally in certain settings derived from Gaussian graphical models.

References

[1] S.A. Andersson, G.G. Wojnar, *Wishart Distributions on Homogeneous Cones*, Journal of Theoretical Probability **17** (2004), 781–818.
[2] S.A. Andersson, G.G. Wojnar, *The Wishart Distributions on Homogeneous Cones*, Acta et Commentationes Universitatis Tartuensis de Mathematica **8** (2004), 3–62.
[3] A.P. Dawid, S.L. Lauritzen, *Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models*, Annals of Statistics **21** (1993), 1272–1317.
[4] G. Letac, H. Massam, *Wishart Distributions for Decomposable Graphs*, Annals of Statistics **35** (2007), 1278–1323.

Statistical challenges in nanoscale biophysics

Samuel Kou

The renowned physicist Richard Feynman once said that "everything that living things can do can be understood in terms of the jigglings and wigglings of atoms." Advances in nanotechnology of the last two decades have brought scientists closer to this "holy grail" than ever before. For the first time scientists were able to study biological processes on an unprecedented nanoscale molecule-by-molecule basis, opening the door to addressing many problems that were inaccessible just a few decades ago.

The new field of nanoscale single-molecule biophysics has attracted much attention from biologists, chemists and biophysicists because nanoscale *single-molecule* experiments offer many advantages over the traditional experiments involving a *population* of molecules. First, by "zooming in" on individual molecules, single-molecule experiments provide data with more accuracy and higher resolution. Second, by isolating, tracking and manipulating individual molecules, single-molecule experiments capture transient intermediates and detailed dynamics of a biological process, the type of information rarely available from traditional population experiments. Third, by following single molecules, scientists can study biological processes directly on the individual molecule level, instead of relying on the extremely difficult task of synchronizing the actions of a population of biomolecules. Fourth, since many important biological functions in a living cell are carried out by single molecules, understanding the behavior of individual biomolecules is a crucial task, for which single-molecule experiments are specifically designed. Many new scientific discoveries have emerged from the nanoscale single-molecule studies.

Advances in nanoscale single-molecule biophysics also bring opportunities and challenges for statisticians and stochastic modelers because of the stochastic nature of single-molecule experiments. First, on the single-molecule level, the laws of statistical and quantum mechanics fundamentally dictate the underlying biological dynamics/processes to be stochastic; their characterization thus requires stochastic models. Second, since the experiments focus on and study only one molecule at a time, the data from single-molecule experiments tend to be much noisier than those from the traditional population experiments because one cannot use the actions of thousands of molecules to average out the noise. Third, in most biophysical experiments, single-molecule experiments in particular, inference of the underlying

stochastic dynamics is usually complicated by the presence of latent processes, which are unobserved but affect the data collection. Fourth, in addition to the preference of analytical tractability, it is important that the stochastic models constructed for biophysical processes should agree with fundamental physical laws and have a sound physical foundation.

In this talk, to illustrate the stochastic modeling and inference problems in the field, we will look at a couple of selected cases, ranging from the utilization of stochastic networks to model single-enzyme reaction dynamics, to likelihood inference of single-molecule fluorescence experiments and to Bayesian data augmentation to handle latent processes.

Adaptive confidence intervals for the test error in classification

Eric B. Laber
(joint work with Susan A. Murphy)

The estimated test error of a learned classification rule is the most commonly reported measure of classifier performance. However, estimating the test error accurately has been established as a nearly hopeless task. Therefore it is crucial that measures of confidence be reported as well. Measures of confidence are typically computed by resampling the estimated test error. However, these approaches do not reliably deliver nominal coverage. We conjecture that the poor performance is partially due to the fact that the test error is a non-smooth functional of the learned classifier. In this article, we present a method for constructing a confidence interval that adapts to amount of non-smoothness in the test error. The proposed method makes no assumptions about the correctness of the model space. We show that the proposed method is consistent under fixed and local alternatives. Moreover, the method provides nominal coverage on a suite of test problems using a range of classification algorithms and sample sizes.

On the sample autocorrelation function of Lévy driven continuous time moving average processes

Alexander Lindner
(joint work with Serge Cohen)

Let $L = (L_t)_{t\in\mathbb{R}}$ be a two-sided Lévy process with expectation 0 and finite variance σ^2, let $f : \mathbb{R} \to \mathbb{R}$ be an L^2-function and let $\mu \in \mathbb{R}$. Then the process $(X_t)_{t\in\mathbb{R}}$ given by

$$X_t = \mu + \int_{\mathbb{R}} f(t-s)\, dL_s, \quad t \in \mathbb{R}, \tag{1}$$

where the integral is defined in the L^2-sense, is called a *continuous time moving average process with mean μ and kernel function f, driven by L*. It is the natural

continuous time analogue of discrete time moving average processes of the form

$$\widetilde{X}_t = \mu + \sum_{i\in\mathbb{Z}} \psi_{t-i} Z_i \tag{2}$$

for a square summable sequence $(\psi_i)_{i\in\mathbb{Z}}$ of coefficients and i.i.d. noise $(Z_i)_{i\in\mathbb{Z}}$.

Now consider the process $(X_t)_{t\in\mathbb{R}}$ as defined in (1) when sampled at integer times $t = 1, 2, 3, \ldots$ For the corresponding sample mean $\overline{X}_n := n^{-1}\sum_{t=1}^n X_t$, it is shown that $\overline{X}_n$ is asymptotically normal as $n \to \infty$ with mean μ and variance $\sigma^2 \int_0^1 (\sum_{j=-\infty}^{\infty} f(u+j))^2\, du$, provided the function $u \mapsto \sum_{j=-\infty}^{\infty} |f(u+j)|$ is in $L^2[0,1]$. Now suppose that $\mu = 0$ and define

$$\gamma_n^*(h) := n^{-1}\sum_{t=1}^n X_t X_{t+h}, \quad \rho_n^*(h) := \frac{\gamma_n^*(h)}{\gamma_n^*(0)}, \quad h \in \mathbb{N}_0,$$

which are specific forms of the sample autocovariance and sample autocorrelation at lag h. Under the condition that L_1 has finite fourth moment and some further conditions on the kernel function f, which in particular assume that the function $u \mapsto \sum_{j=-\infty}^{\infty} f(u+j)^2$ is in $L^2[0,1]$ and that the sequence $(\gamma(k) = \mathrm{Cov}(X_0, X_k))_{k\in\mathbb{Z}}$ of autocovariance functions is square summable, it is shown that $(\gamma_n^*(0), \ldots, \gamma_n^*(h))$ and $(\rho_n^*(1), \ldots, \rho_n^*(h))$ are asymptotically normal with a certain asymptotic variance. The elements of the asymptotic covariance matrix of $(\rho_n^*(1), \ldots, \rho_n^*(h))$ are of the form

$$w_{ij} = \widetilde{w}_{ij} + v_{ij},$$

where $\widetilde{w}_{ij}$ is given by Bartlett's formula (cf. Brockwell and Davis [1], Theorem 7.2.1), and v_{ij} is an additional term which depends on the fourth moment of L_1 and further properties of the kernel function f. This is in sharp contrast to the well known asymptotic behaviour of the sample autocorrelation function of discrete time moving average processes with i.i.d. noise as in (2), where this extra term does not appear (e.g. Brockwell and Davis [1], Theorem 7.2.1).

The results can be applied to show asymptotic normality of a moment estimator of the Hurst index of fractional Lévy processes. For this, let L be a Lévy process with finite variance and expectation zero and consider for $d \in (0, 1/2)$ the fractional Lévy process $M_d(t) := \frac{1}{\Gamma(d+1)} \int_{-\infty}^{\infty} [t-s)_+^d - (-s)_+^d]\, dL_s$ as defined in Marquardt [3]. The corresponding fractional Lévy noise based on increments of length 1 is given by

$$X_t = M_d(t) - M_d(t-1) = \frac{1}{\Gamma(d+1)} \int_{-\infty}^{\infty} [(t-s)_+^d - (t-s-1)_+^d]\, dL_s.$$

An application of the results of the asymptotic normality of the autocorrelation function for continuous time moving average processes then shows that the moment estimator

$$\widehat{H}_n := \frac{1}{2}\left(1 + \frac{\log(\rho^*(1)+1)}{\log 2}\right)$$

for the Hurst parameter $H := d + 1/2$ is asymptotically normal if $d \in (0, 1/4)$ and L_1 has finite fourth moment. Since fractional Lévy noise is shown to be generally

not strongly mixing, this cannot be deduced from standard results on strongly mixing time series. The moment estimator is however strongly consistent for each $d \in (0, 1/2)$, since the fractional Lévy noise is mixing in the ergodic theoretic sense. An asymptotically normal estimator for general $d \in (0, 1/2)$ can also be obtained by applying the results to differenced fractional Lévy noise.

The results of the talk are based on [2].

References

[1] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods*, 2nd Edition. Springer (1991).

[2] S. Cohen and A. Lindner, *Asymptotic behaviour of the sample autocovariance function of Lévy driven continuous time moving average processes* (2009), in preparation.

[3] T. Marquardt, *Fractional Lévy processes with an application to long memory moving average processes*, Bernoulli **12**, 1099–1126 (2006).

Proving consistency of non-standard kernel estimators

David M. Mason

Here is our basic setup. Let $X, X_1, X_2, \ldots,$ be i.i.d. random variables from a probability space $(\Omega, \mathcal{A}, P)$ to a measure space $(S, \mathcal{S})$, and let $\mathcal{G}$ denote a class of measurable real valued functions

$$g \text{ of } (u, h) \in S \times (0, 1].$$

We study classes of statistics of the following form: For any $n \geq 1$, $g \in \mathcal{G}$ and $0 < h \leq 1$ define,

$$g_{n,h} := n^{-1} \sum_{i=1}^{n} g\left(X_i, h\right).$$

We describe a general result that says that under suitable regularity conditions, with probability one,

$$\limsup_{n\to\infty} \sup_{a_n \leq h \leq h_0} \sup_{g\in\mathcal{G}} \frac{\sqrt{n}\,|g_{n,h} - \mathbb{E}g_{n,h}|}{\sqrt{h\left(|\log h| \vee \log\log n\right)}} < \infty.$$

Numerous applications of this result to function estimation are given.

This talk is partially based upon joint work with Julia Dony, Uwe Einmahl and Jan Swanepoel. The list of references given below contains papers relevant to our talk as well as the large sample theory facts used to prove our results.

References

[1] J. Dony, U. Einmahl and D.M. Mason, *Uniform in bandwidth consistency of local polynomial regression function estimators*, Austrian Journal of Statist. **35** (2006), 105–120.

[2] U. Einmahl and D.M. Mason, *An empirical process approach to the uniform consistency of kernel-type function estimators*, J. Theoretical Prob. **13** (2000), 1–37.

[3] U. Einmahl and D.M. Mason, *Uniform in bandwidth consistency of kernel-type function estimators*, Ann. Statist. **33** (2005), 1380–1403.

[4] D.M. Mason and J. Swanepoel, *A general result on the uniform in bandwidth consistency of kernel-type function estimators*, submitted for publication.

[5] D. Nolan and J.S. Marron, *Uniform consistency of automatic and location-adaptive delta–sequence estimators*, Probab. Th. Rel. Fields **80** (1989), 619–632.
[6] M. Talagrand, *Sharper bounds for Gaussian and empirical processes*, Ann. Probab. **22** (1994), 28–76.
[7] M. Talagrand, M., *New concentration inequalities in product spaces*, Invent. Math. **126** (1996), 505–563.
[8] A.W van der Vaart and J.A. Wellner, *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

The extremogram – a correlogram for extremal events

Thomas Mikosch
(joint work with R.A. Davis)

We consider a strictly stationary sequence of random vectors whose finite-dimensional distributions are jointly regularly varying (regvar) with some positive index. This class of processes includes among others ARMA processes with regvar noise, GARCH processes with normal or student noise, and stochastic volatility models with regvar multiplicative noise. We define an analog of the autocorrelation function, the extremogram, which only depends on the extreme values in the sequence. We also propose a natural estimator for the extremogram and study its asymptotic properties under α-mixing. We show asymptotic normality, calculate the extremogram for various examples and consider spectral analysis related to the extremogram. Finally, we propose to use the stationary bootstrap to generate confidence bounds.

The paper is available under www.math.ku.dk/∼mikosch and will appear in Bernoulli.

Bayesian inference for D-vine pair-copula constructions: Estimation and model selection

Aleksey Min
(joint work with Claudia Czado)

Copulas are nowadays a standard tool for stochastic modeling in different fields of applied science. Therefore the construction of flexible multivariate copulas as well as the right choice of a copula family, i.e. model selection for copulas, have become extremely important. Recently Aas, Czado, Frigessi and Bakken (2009) have advocated pair-copula constructions (PCC) which have been found as a most successful way of construction of multivariate copulas in many empirical studies. In this talk we first discuss Bayesian estimation in PCC's. Further using pair-copula constructions we approach the model selection problem for copulas to identify (conditional) independence, present in data, in a fully Bayesian framework. For this problem we derive and implement a reversible jump Markov chain Monte Carlo (RJ MCMC) algorithm. Building blocks of PCC's are fixed as bivariate t-copulas. However the methodology is general and can easily be extended to all

known bivariate copula families. Our approach with the RJ MCMC solves model selection and estimation problems for PCC's simultaneously. The effectiveness of the developed algorithms is shown in simulations and their usefulness is illustrated in a real data application.

References

[1] K. Aas, C. Czado, A. Frigessi and H. Bakken, *Pair-copula constructions of multiple dependence*, Insurance Mathematics and Economics **44** (2009), 182-198.

[2] A. Min and C. Czado, *Bayesian inference for multivariate copulas using pair-copula constructions*, Preprint (2008) available at *http://www-m4.ma.tum.de/Papers/Czado/cc-pubs.html*.

[3] A. Min and C. Czado, *Bayesian model selection for multivariate copulas using pair-copula constructions*, Preprint (2009) available at *http://www-m4.ma.tum.de/Papers/Czado/cc-pubs.html*.

Primal and Dual Formulations in Density Estimation: Some Theoretical Consequences

Ivan Mizera

In the series of papers [3, 4, 5, 6], co-authored with Roger Koenker, we explored dual formulations (in the sense of convex conjugacy) to nonparametric maximum likelihood density estimation, both in penalized and shape-constrained settings. All the cases we considered fall under a general scheme

$$\text{(P)} \qquad -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \lambda J(Dg) + \int_\Omega \psi(g(x))\,dx \;\rightarrowtail\; \inf_g !$$

where X_i denote the observed datapoints, D stands for a differential operator, the kind that appear in typical regularization formulations (for instance, the first, second, or third derivative of a univariate function; the Hessian of a bivariate function), J is a convex function, and ψ is a convex, nondecreasing real function.

If J is an integral L^1-norm, or square of the L^2-norm, the problem (P) represents the classical formulation of a penalized maximum likelihood density estimation—with tuning parameter $\lambda > 0$. The L^2 choice, for instance, together with D representing the operator of the third derivative and $\psi(z) = \mathrm{e}^z$, yields the classical proposal [8]; the L^1 choice, with the same ψ, and D corresponding to the second derivative yields the estimator proposed in [3].

The general scheme (P), however, also allows for the accommodation of various shape-constrained density estimation prescriptions: for example, keeping the same ψ, and the same, second derivative operator for D, but taking the indicator function of the set $(-\infty, 0]$ (the indicator function of a set E is defined to assume 0 for all $x \in E$, and $+\infty$ otherwise) for J makes the parameter λ irrelevant, and (P) then defines one of the log-concave density estimates studied by [9, 7, 2].

Elaborating on the fact that the "primal" (P) is a convex problem, we derive its conjugate dual formulation

$$\text{(D)} \qquad -\int_\Omega \psi^*(P_n - D^*u)\,dx - \lambda J^*\Big(\frac{u}{\lambda}\Big) \looparrowright \max_u!$$

where P_n is the empirical probability supported by the datapoints, ψ^* and J^* denote functions conjugate to ψ and J, respectively, and D^* stands for the operator adjoint to D (the tuning parameter λ stays the same). For suitable ψ, (D) can be interpreted as a specific forms of maximum entropy formulation—as is apparent for $\psi(z) = \mathrm{e}^z$, where $\psi^*(u) = u\log u - u$ and (D) then maximizes the Shannon entropy of $P_n - D^*u$, the quantity that turns out to be equal to the estimated density f (unlike g in (P), which is its transformations, its form following from the specification of ψ).

Formulations (P) and (D) were in [3, 4, 5] considered mostly in the discretized setting, and utilized there rather for immediate, data-analytic objectives—like more efficient computation [3], or density estimation under more flexible shape constraints [6]; of interest was also the connection to the so-called taut string density estimation method [1], explored in [5]. Our focus now are exact duality results in the functional, *infinite-dimensional* setting—in particular, our objective is to establish the strong duality in this context, a mathematical property that ensures that the minimal value of the primal is the maximal of the dual—the fact that consequently links primal and dual formulations via so-called extremal relations, in this particular case stipulating that $f = P_n - D^*u = \psi'(g)$, where u and g are the solutions of dual and primal, respectively.

We explore theoretical consequences of duality results cast in the functional setting—those relevant for the choice of the domain for density estimation, and for the moment-preservation properties of the resulting estimates. In particular, we obtain that—unlike in classical smoothing spline fitting, where increasing dimension means adding derivatives of higher orders into the regularization penalty—in density estimation derivatives of order three are sufficient to keep the problem well-posed, in arbitrary dimension. This, among other things, vindicates the choice of third derivative in [8].

References

[1] P. L. Davies and A. Kovac, *Densities, spectral densities and modality*, Ann. Statist., **32** (2004), 1093–1136.

[2] L. Dümbgen and K. Rufibach, textitMaximum likelihood estimation of a log-concave density: Basic properties and uniform consistency, Bernoulli **15** (2009), 40–68.

[3] R. Koenker and I. Mizera, *Density estimation by total variation regularization*, Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum (Vijay Nair, ed.), World Scientific, 2007, 613–634.

[4] R. Koenker and I. Mizera, *The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and stretched strings*, Prague Stochastics 2006, Proceedings of the joint session of 7th Prague Symposium on Asymptotic Statistics and 15th Prague conference on Information Theory, Statistical Decision Functions and Random Processes, held in Prague from August 21 to 25, 2006 (M. Hušková and M. Janžura, eds.), Prague, Matfyzpress, 2006, 145–157.

[5] R. Koenker and I. Mizera, *Primal and dual formulations relevant for the numerical estimation of a probability density via regularization*, Tatra Mountains Mathem. Publications **39** (2008), 255–264.
[6] R. Koenker and I. Mizera, *Quasi-concave density estimation*, submitted.
[7] J. K. Pal, M. Woodroofe, and M. Meyer, *Estimating a Polya frequency function*, in Complex Datasets and Inverse Problems: Tomography, Networks and Beyond (R. Liu, W. Strawderman, and C.-H. Zhang, eds.), IMS Lecture Notes-Monograph Series 54, 239–249.
[8] B. W. Silverman, *On the estimation of a probability density function by the maximum penalized likelihood method*, Ann. Statist. **10** (1982), 795–810.
[9] G. Walther, *Detecting the presence of mixing with multiscale maximum likelihood*, J. Amer. Statist. Assoc. **97** (2002), 508–513.

Statistical Aspects of COGARCH Modelling

Gernot Müller

(joint work with Robert Durand, Jean Jacod, Claudia Klüppelberg, Ross Maller)

In this talk we discuss estimation and testing for the COGARCH model. The COGARCH model was introduced by Klüppelberg et al. (2004) as a continuous time version of the discrete time GARCH model and is constructed directly from a single univariate background driving Lévy process L. The model can be defined by the two stochastic differential equations

$$dG_t = \sigma_{t-} dL_t \tag{1}$$

$$d\sigma_t^2 = \beta\, dt - \eta\sigma_{t-}^2\, dt + \varphi\sigma_{t-}^2\, d[L,L]_t^{(d)} \tag{2}$$

where G is the integrated COGARCH process and σ^2 the variance process. The three parameters $\beta > 0$, $\eta > 0$, and $\varphi \geq 0$ are assumed to be unknown. Note that the Lévy process L occurs both in Equations (1) and (2), the variance process, however, is only affected by the discrete part of the quadratic variation of L. In Maller et al. (2008) it has been shown that it is possible to approximate the COGARCH with an embedded sequence of discrete time GARCH models which converges to the continuous time model in a strong sense (in probability, in the Skorokhod metric), as the discrete approximating grid grows finer. One can use this result to fit the COGARCH to irregularly spaced time series data.

As an application we confirm Merton's hypothesis on the risk-return tradeoff using COGARCH. We find clear evidence of a positive relationship between return and risk in daily data covering the period from 1953 to 2007, thus providing empirical verification of Merton's theorised relationship. Our model estimates a highly significantly positive risk premium of about 8% p.a., consistent with other published estimates such as those of Lundblad (2007) who found a significant relationship using monthly data from 1836 to 2003, but not in monthly data for the period 1950 to 2003; we conclude in favour of Merton's theory also for the period after 1950. As a sidelight, our COGARCH model estimates that, over a long period, the weekend is equivalent, in terms of volatility, to about 0.3-0.5 regular trading days. This part of the talk is joint work with Robert Durand and Ross Maller.

Furthermore, we discuss an asymptotic test for the special feature of COGARCH, that there is a fixed functional relationship between the jumps of the COGARCH process and the jumps of the corresponding volatility process. The test only investigates the bigger jumps of the COGARCH (which is, in practice, identified with a log-price process) and uses local volatility estimates calculated from observation windows before and after the jumps under consideration. The null of the test is the fixed relationship conditional on the fact that there is at least one relevant (i.e. sufficiently big) jump. We apply the test to high-frequency data from the S&P 500. More precisely, we look at 5 minutes log-returns for the ten years 1998 to 2007 separately. Using time windows of one hour forward and one hour backward to estimate the local volatility, we reject the null on a 10% level only for one of the ten years (for 2005), and even never reject it on a 5% level. This latter part of the talk is joint work with Jean Jacod and Claudia Klüppelberg.

References

[1] Durand, R. B., Maller, R. A., Müller, G., *The Risk Return Tradeoff: A COGARCH Analysis in Favour of Merton's Hypothesis*, Preprint, University of Western Australia, Australian National University and Technische Universität München (2009).
[2] Jacod, J., Klüppelberg, C., Müller, G., *Testing for COGARCH*, Working paper, Université Paris VI and Technische Universität München (2009).
[3] Klüppelberg, C., Lindner, A., Maller, R. A., *A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour*, Journal of Applied Probability **41** (2004), 601–622.
[4] Lundblad, Ch., *The risk return tradeoff in the long run: 1836–2003*, Journal of Financial Economics **85** (2007), 123–150.
[5] Maller, R. A., Müller, G., Szimayer, A., *GARCH modelling in continuous time for irregularly spaced time series data*, Bernoulli **14** (2008), 519–542.

Towards Brain Computer Interfacing

Klaus-Robert Müller
(joint work with Benjamin Blankertz, Gabriel Curio)

We outline the Berlin Brain-Computer Interface (BBCI), a system which enables us to translate brain signals from movements or movement intentions into control commands [5]. The main contribution of the BBCI, which is a non-invasive EEG(electroencephalography)-based BCI system, is the first time use of advanced data analysis methods from statistics, machine learning and signal processing [6, 7, 5, 9]. These techniques allow to adapt to the specific brain signatures of each user with literally no training. In BBCI a calibration session of about 20min is necessary to provide a data basis from which the individualized brain signatures are inferred (cf. [2, 9, 5]). This is very much in contrast to conventional BCI approaches that rely on operand conditioning and need extensive subject training of the order 50-100 hours (cf. [5]). Our machine learning concept thus allows to achieve high quality feedback already after the very first session [2].

This talk has reviewed a broad range of investigations and experiments that have been performed within the BBCI project. Physiologically the BBCI decodes voluntary modulations of sensorimotor rhythms caused by motor imagery (left hand vs. right hand vs. foot) which can be readily translated into a continuous feedback signal.

In addition to these general paradigmatic BCI results, we provided a condensed outline of the underlying machine learning and signal processing techniques that make the BBCI succeed: after artifact removal, a subject specific frequency prefiltering is performed, the multivariate signal (64 channels at 200hz) is smoothed in the spatial and spectral domain using the so-called common spatial pattern approach [3]. Depending on the subject additional to these resulting features dynamic components (such as channel-wise AR model fits) or slow variations (cf. [5]) are used finally resulting in up to 1000 dimensional input features that are subsequently to be classified in real-time. Note that the available data that is acquired during the calibration session contains only 200 data points (large p, small N problem). While the classifiers in use in BBCI are typically linear in nature [6, 7, 5, 9], it depends strongly on the subject and the physiological paradigm whether a nonlinear method may increase performance. In any case heavy regularisation and robust (e.g. L_1 optimizing) methods are mandatory for successful performance (cf. [7, 9]). An important challenge that is commonly found in EEG-BCI data are effects of non-stationarity: from the calibration session to the feedback session the underlying probability distribution is subject to change. A possible remedy against this aspect which is rather common for general real-world data are models that take into account covariate shifts [8], invariant representations or projection methods that consider only the stationary subspaces of the high dimensional EEG data stream [4].

Results of a recent feedback study with 6 healthy subjects with no or very little experience with BCI control: half of the subjects achieved an information transfer rate above 35 bits per minute (bpm). Furthermore one subject used the BBCI to operate a mental typewriter in free spelling mode. The overall spelling speed was 4.5-8 letters per minute including the time needed for the correction of errors. This opens up a large number of possible applications both in rehabilitation as well as general man machine interaction for our novel BCI technology.

For further, more detailed information please refer to `www.bbci.de` where also a large repository of high quality data is available for further study [1].

References

[1] Blankertz, B., Müller, K.-R., Krusienski, D.J., Wolpaw, J.R., Schlögl, A., Pfurtscheller, G., Millan, J., Schröder, M., Birbaumer, N., *The BCI competition III: Validating Alternative Approaches to Actual BCI Problems*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **14(2)** (2006), 153-159.

[2] Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Curio, G., *The non-invasive Berlin Brain-Computer Interface: Fast Acquisition of Effective Performance in Untrained Subjects*, NeuroImage **37(2)** (2007), 539-550.

[3] Blankertz, B., Tomioka, R., Lemm , S., Kawanabe, M., Müller, K.-R., *Optimizing Spatial Filters for Robust EEG Single-Trial Analysis*, IEEE Signal Processing Magazine, **25(1)** (2008), 41-56.
[4] von Bünau, P., Meinecke, F.C., Kiraly, F. and Müller, K.-R., Estimating the Stationary Subspace from Superimposed Signals. *Physical Review Letters*, 103, 214101, 2009.
[5] Dornhege, G., Millan, J., Hinterberger, T., McFarland, D., Müller, K.-R. (eds.), *Toward Brain Computer Interfacing*, MIT Press (2007)
[6] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., *An Introduction to Kernel-Based Learning Algorithms*, IEEE Transactions on Neural Networks **2(2)** (2001), 181-201.
[7] Müller, K.-R., Anderson, C., Birch, B., *Linear and nonlinear methods for Brain Computer Interfaces*, IEEE Transactions on Neural Systems and Rehabilitation Engineering **11(2)** (2003), 165-169.
[8] Sugiyama, M., Krauledat, M., Müller, K.-R., *Covariate Shift Adaptation by importance weighted cross validation*, Journal of Machine Learning Research **8** (2007), 985-1005.
[9] Tomioka, R., Müller, K.-R., A regularized discriminative framework for EEG based communication. *Neuroimage*, 49(1), 415–432, 2010.

The Estimation of Different Scales in Microstructure Noise Models from a Nonparametric Regression Perspective

Axel Munk and Johannes Schmidt-Hieber
(joint work with T. Tony Cai)

Introduction and Model. Consider the models

$$Y_{i,n} = \int_0^{i/n} \sigma(s)\, dW_s + \tau\left(\frac{i}{n}\right)\epsilon_{i,n}, \quad i = 1, \ldots, n, \tag{1}$$

and

$$\tilde{Y}_{i,n} = \sigma\left(\frac{i}{n}\right) W_{i/n} + \tau\left(\frac{i}{n}\right)\epsilon_{i,n}, \quad i = 1, \ldots, n, \tag{2}$$

respectively, where $(W_t)_{t\geq 0}$ denotes a Brownian motion and $\epsilon_{i,n}$ is so called microstructure noise, i.e. we assume $\epsilon_{i,n}$ i.i.d., $\mathrm{E}\left(\epsilon_{i,n}^2\right) = 1$ and $\mathrm{E}\left(\epsilon_{i,n}^4\right) < \infty$. $(W_t)_{t\geq 0}$ and $(\epsilon_{1,n}, \ldots, \epsilon_{n,n})$ are assumed to be independent, and σ and τ are unknown, positive and deterministic functions. The problem is to estimate the scale functions σ and τ, pointwise. Model (1) appears in a more general form in modelling log-returns on frequencies up to one second (see [3]). The measurement error is induced by market frictions, such as bid-ask spreads and rounding. In this model σ is the volatility of the asset and is the quantity we are mainly interested in whereas τ is the noise level and is sometimes considered as a quality measure of a market.

Estimation of instantaneous volatility and lower bounds.. In a first step we investigate the case where σ and τ are constants and $\epsilon_{i,n} \sim \mathcal{N}(0,1)$. Then models (1) and (2) reduce to

$$Y_{i,n} = \sigma W_{i/n} + \tau\epsilon_{i,n}, \quad i = 1, \ldots, n.$$

Since $Y = (Y_{1,n}, \ldots, Y_{n,n})'$ is multivariate centered Gaussian it is described completely by its covariance $\mathrm{Cov}(Y) = \sigma^2 K_n + \tau^2 I$, where $K_n = (i/n \wedge j/n)_{i,j=1,\ldots,n}$

and I_n denotes the $n \times n$ identity matrix. Let $\lambda_{i,n}$ denote the eigenvalues of K_n then there is an orthogonal transformation D_n not depending on the unknown quantities, such that the transformed observation vector $Z = D_n Y$ has independent components $Z_i \sim \mathcal{N}\left(0, \sigma^2\lambda_i + \tau^2\right)$, where $\lambda_i \sim n/i^2$. (see [1]). Adopting the statistical inverse problem point of view, this suggests that the number of observation in the spectral domain usable for estimation of σ and τ is of order $\sqrt{n}$ and n, respectively. This suggests $n^{1/4}$ and $n^{1/2}$ as the optimal rates of convergence. To be precise, we can construct estimators $\hat{\tau}^2$ and $\hat{\sigma}^2$ that are minimax sharp, i.e. they reach the following bound (see [2] and [1])

Theorem 1. *For any* $\epsilon > 0$

$$\lim_{n\to\infty} \inf_{\hat{\tau}^2} \sup_{\sigma,\tau>\epsilon} (\sigma\tau)^{-4} \left(\mathrm{E}\left(n \left(\hat{\tau}^2 - \tau^2 \right)^2 \right) - 2\tau^4 \right) = 0,$$

$$\lim_{n\to\infty} \inf_{\hat{\sigma}^2} \sup_{\tau,\sigma>\epsilon} (\sigma\tau)^{-8} \left(\mathrm{E}\left(n^{1/2} (\hat{\sigma}^2 - \sigma^2)^2 \right) - 8\tau\sigma^3 \right) = 0.$$

Construction of a sharp estimator of τ^2 is easy, for estimation of σ^2 we introduce a splitting technique of the spectral observations (see [1]).
In the second part of the talk we extend this technique to the general models (1) and (2). In a first step we construct $n^{1/4}$ consistent estimators for $\langle \sigma^2, \phi_k \rangle$ with respect to the particular basis system $\{\psi_k, k = 0, \ldots\} := \left\{1, \sqrt{2}\cos(k\pi t), k = 1, \ldots\right\}$. Further, we introduce the corresponding truncated Sobolev ellipsoid $\Theta_s^b(\alpha, C)$ as the space of all L_2-functions bounded by universal constants $0 < C_1, C_2 < \infty$ from below and above, such that the series coefficients $(\theta_k)_{k\geq 0}$ with respect to the basis system above, satisfy $\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \leq C$.

Theorem 2. *Suppose* $Q, \bar{Q} > 0$ *are fixed constants and let* $\beta > 5/4$. *Assume model (1) and* $\alpha > 3/4$ *or model (2) and* $\alpha > 3/2$. *Then it holds for* $N^* = n^{1/(4\alpha+2)}$

$$\sup_{\tau^2 \in \Theta_s^b\left(\beta,\bar{Q}\right),\ \sigma^2 \in \Theta_s^b(\alpha,Q)} \mathrm{MISE}\left(\hat{\sigma}^2_{N^*}\right) = O\left(n^{-\alpha/(2\alpha+1)}\right).$$

The following lower bound reveals the estimator as rate minimax.

Theorem 3. *Assume model (1) or model (2),* $\alpha \in \mathbb{N}^*$ *and* $\tau > 0$. *Then there exists a* $C > 0$, *such that*

$$\liminf_{n\to\infty} \inf_{\hat{\sigma}^2_n} \sup_{\sigma^2 \in \Theta_s^b(\alpha,Q)} \mathrm{E}\left(n^{\frac{\alpha}{2\alpha+1}} \left\| \hat{\sigma}^2 - \sigma^2 \right\|_2^2 \right) \geq C.$$

Hence this shows that $n^{\alpha/(4\alpha+2)}$ is the optimal rate of convergence, which is"half" of the usual rates obtained in nonparametric regression. This is due to the additional degree of ill-posedness induced by microstructure noise. The key step in the proof is a new bound on Kullback-Leibler divergence of two multivariate normal vectors ([4]).

Computation and Simulation. Monte Carlo simulations show that the proposed estimator of the instantaneous (spot) volatility works quite well, even when the normality assumption of the microstructure noise does not hold and heavy tails

are present. Furthermore, for instantaneous volatility which is non deterministic our simulation shows that the estimator still captures the major features, such as aprupt changes in size, quite reasonably.

References

[1] T. Cai, A. Munk, and J. Schmidt-Hieber, (2009), *Sharp minimax estimation of the variance of Brownian motion corrupted by Gaussian noise*, Stat. Sinica, to appear.
[2] A. Gloter and J. Jacod (2001), *Diffusions with measurement errors. I. Local Asymptotic Normality*, ESAIM: Probability and Statistics **5**, 225–242.
[3] J. Jacod, Y. Li, P. Mykland, M. Podolskij, and M. Vetter (2009), *Microstructure noise in the continuous case: The pre-averaging approach*, Stoch. Proc. and their Appl. **119**, 2249–2276.
[4] A. Munk and J. Schmidt-Hieber (2009), *Nonparametric estimation of the volatility function in a high-frequency model corrupted by noise*, Available on arxiv.org.

Generalized Binary Search

Robert Nowak

Consider statistical learning problems of the following form. Consider a finite, but potentially very large, collection of binary-valued functions H defined on a domain X (e.g., an ϵ-cover of an uncountable class, with respect to a measure on X). H will be called the *hypothesis space* and X will be called the *query space*. Each $h \in H$ is a mapping from X to $\{-1, 1\}$. Assume that the functions in H are unique and that one function, $h^* \in H$, produces the correct binary labeling. For each query $x \in X$, the value $h^*(x)$, corrupted with independently distributed binary noise, is observed. If the queries are drawn randomly, then this leads to the standard binary classification problem. Here we assume that we have control over the selection of queries. The queries can be selected in a sequential fashion, using past information to guide the selection, and the goal is to determine h^* through as few queries from X as possible. If the queries were noiseless, then they are usually called *membership queries* to distinguish them from other types of queries [2]; here we will simply refer to them as queries. Problems of this nature arise in many applications , including channel coding [8, 17], disease diagnosis [14], job scheduling [13], image processing [12, 11], computer vision [16, 6], computational geometry [1], and active learning [4, 3, 15].

Past work has provided a partial characterization of this problem. If the responses to queries are noiseless, then selecting the optimal sequence of queries from X is equivalent to determining an optimal binary decision tree, where a sequence of queries defines a path from the root of the tree (corresponding to H) to a leaf (corresponding to a single element of H). In general the determination of the optimal tree is NP-complete [10]. However, there exists a greedy procedure that yields query sequences that are within an $O(\log |H|)$ factor of the optimal search tree depth [5, 13, 14, 1, 4], where $|H|$ denotes the cardinality of H. The greedy procedure is referred to as *Generalized Binary Search* (GBS) [4, 15] or the *splitting algorithm* [13, 14, 5]), and it reduces to classic binary search in special cases [15]. The GBS algorithm is outlined in Figure . At each step GBS selects a query

that results in the most even split of the hypotheses under consideration into two subsets responding $+1$ and -1, respectively, to the query. The correct response to the query eliminates one of these two subsets from further consideration. Since the hypotheses are assumed to be distinct, it is clear that GBS terminates in at most $|H|$ queries (since it is always possible to find query that eliminates at least one hypothesis at each step). In fact, there are simple examples demonstrating that this is the best one can hope to do in general [13, 14, 5, 4, 15]. However, it is also true that in many cases the performance of GBS can be much better [1, 15]. In general, the number of queries required can be bounded in terms of a combinatorial parameter of H called the extended teaching dimension [2, 7] (also see [9] for related work). Alternatively, there exists a geometric relation between the pair (X, H), called the *neighborly* condition, that is sufficient to bound the number of queries needed [15]. The number of queries an algorithm requires to confidently identify h^* is called the *sample complexity* of the algorithm. Under the neighborly condition, the sample complexity of GBS is optimal. We also present a noise-tolerant version of GBS to handle errors. The noise-tolerant GBS algorithm also achieves the optimal sample complexity, and we are not aware of any other algorithm with this capability.

Generalized Binary Search (GBS)
initialize: $n = 0$, $H_0 = H$.
while $|H_n| > 1$
1) Select $x_n = \arg\min_{x \in X} |\sum_{h \in H_n} h(x)|$.
2) Query with x_n to obtain response $y_n = h^*(x_n)$.
3) Set $H_{n+1} = \{h \in H_n : h(x_n) = y_n\}$, $n = n + 1$.

FIGURE 1. Generalized Binary Search algorithm.

REFERENCES

[1] E. M. Arkin, H. Meijer, J. S. B. Mitchell, D. Rappaport, and S.S. Skiena. Decision trees for geometric models. *Intl. J. Computational Geometry and Applications*, 8(3):343–363, 1998.

[2] D. Angluin. Queries revisited. *Springer Lecture Notes in Comp. Sci.: Algorithmic Learning Theory*, pages 12–31, 2001.

[3] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Conf. on Learning Theory (COLT)*, 2007.

[4] S. Dasgupta. Analysis of a greedy active learning strategy. In *Neural Information Processing Systems*, 2004.

[5] M. R. Garey and R. L. Graham. Performance bounds on the splitting algorithm for binary testing. *Acta Inf.*, 3:347–355, 1974.

[6] D. Geman and B. Jedynak. An active testing model for tracking roads in satellite images. *IEEE Trans. PAMI*, 18(1):1–14, 1996.

[7] T. Hegedüs. Generalized teaching dimensions and the query complexity of learning. In *8th Annual Conference on Computational Learning Theory*, pages 108–117, 1995.

[8] M. Horstein. Sequential decoding using noiseless feedback. *IEEE Trans. Info. Theory*, 9(3):136–143, 1963.

[9] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins. How many queries are needed to learn? *J. ACM*, 43(5):840–862, 1996.
[10] L. Hyafil and R. L. Rivest. Constructing optimal binary decision trees is NP-complete. *Inf. Process. Lett.*, 5:15–17, 1976.
[11] A. P. Korostelev and J.-C. Kim. Rates of convergence fo the sup-norm risk in image models under sequential designs. *Statistics & Probability Letters*, 46:391–399, 2000.
[12] A. P. Korostelev. On minimax rates of convergence in image models under sequential design. *Statistics & Probability Letters*, 43:369–375, 1999.
[13] S. R. Kosaraju, T. M. Przytycka, and R. Borgstrom. On an optimal split tree problem. *Lecture Notes in Computer Science: Algorithms and Data Structures*, 1663:157–168, 1999.
[14] D. W. Loveland. Performance bounds for binary testing with arbitrary weights. *Acta Informatica*, 22:101–114, 1985.
[15] R. Nowak. Generalized binary search. In *Proceedings of the Allerton Conference, Monticello, IL*, (www.ece.wisc.edu/∼nowak/gbs.pdf) 2008.
[16] M.J. Swain and M.A. Stricker. Promising directions in active vision. *Int. J. Computer Vision*, 11(2):109–126, 1993.
[17] K. Sh. Zigangirov. Upper bounds for the error probability of feedback channels. *Probl. Peredachi Inform.*, 6(2):87–82, 1970.

Statistical challenges in the analysis of neuroscience data

Liam Paninski
(joint work with Yashar Ahmadian, Yuriy Mischchenko, Joshua Vogelstein)

Our primary research focus is the analysis of neural data. A number of very challenging high-dimensional problems arise in this field; here we summarize three problems on which we have made some progress recently.

1) *Inference of connectivity in large neuronal networks given limited noisy observations.* It has recently become possible to record simultaneously from multiple neurons in real neuronal networks (not to be confused with artificial neural networks), though in many cases the available observations are still quite noisy. A number of major open questions in neuroscience involve the connectivity in large neuronal networks. However, reconstructing the connectivity from the available noisy, incomplete data remains a challenging open problem. We have developed methods for inferring the connectivity from large-scale recordings, based on a generalized linear model and sequential Monte Carlo framework [3], and are currently testing these methods on real data.

2) *Optimal filtering and smoothing of high-dimensional voltage signals on dendritic trees.* Many neurons have highly articulated and geometrically-complex structures known as "dendrites" which play an important role in neuronal communication and computation. Optimal filtering of noisy voltage signals on these dendritic trees is a key problem in cellular neuroscience. However, the state variable in this problem — the vector of voltages at every compartment on the tree — is very high-dimensional: typical realistic multicompartmental models have on the order of $N = 10^4$ degrees of freedom. Standard implementations of the Kalman filter require $O(N^3)$ time and $O(N^2)$ space, and are therefore impractical. However, it is possible to take advantage of three special features of the dendritic filtering

problem to construct an efficient filter [1]: (1) dendritic dynamics are governed by a cable equation on a tree, which may be solved using sparse matrix methods in $O(N)$ time; and current methods for observing dendritic voltage (2) provide low SNR observations and (3) only image a few compartments (< 100 or so) at a time. The idea is to approximate the Kalman equations in terms of a low-rank perturbation of the steady-state (zero-SNR) solution, which may be obtained in $O(N)$ time using junction-tree methods that exploit the sparse tree structure of dendritic dynamics. The resulting methods give a very good approximation to the exact Kalman solution, but only require $O(N)$ time and space.

3) *Decoding visual imagery and movies given the responses of large populations of retinal ganglion cells.* A third major problem in systems neuroscience involves decoding the information encoded in the spiking activity of large neural populations. The visual system presents a number of interesting challenges, since the information encoded by the eye (time-varying visual movies) is very high-dimensional. We have developed fast MAP decoding techniques that make heavy use of ideas from convex optimization and numerical linear algebra [2]. We are currently analyzing these decoded movies to better understand what information the retina sends to the brain about the visual world, and what information is discarded.

References

[1] Paninski, L. *Fast Kalman filtering on quasilinear dendritic trees*, Under review (2009).
[2] Pillow, J., Ahmadian, Y. & Paninski, L. (2009). *Model-based decoding, information estimation, and change-point detection in multi-neuron spike trains*, Neural Computation, Under review (2009).
[3] Vogelstein, J., Watson, B., Packer, A., Yuste, R., Jedynak, B. & Paninski, L. *Spike inference from calcium imaging using sequential Monte Carlo methods*, Biophysical Journal, In press (2009).

Application of the Malliavin calculus to statistical problems on Gaussian fields

Mark Podolskij

In this talk we present some recent techniques for proving central limit theorems using standard operators of Malliavin calculus and apply them to various statistical problems.

Quite often people use (discrete or continuous time) Gaussian processes to generate models for certain phenomena in science. However, when it comes to estimation problems in those models (especially for high frequency data) there is a need for general central limit theorems, because Gaussian process usually do not have a nice probabilistic structure (they are neither semimartingales nor Markov processes).

Let us briefly describe some examples which we have in mind:

Problem 1: Let $(X_i)_{i\geq 1}$ be a stationary sequence of random variables with $X_i \sim N(0,1)$ and $r_k = \mathrm{cov}(X_1, X_{1+k})$. Under which assumptions on r_k do we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) \Rightarrow N(0, \sigma^2)$$

if $Eg(X_1) = 0$?

Problem 2: Let $(G_t)_{t\geq 0}$ be a Gaussian process with centered and stationary increments. What is the weak limit of the functionals

$$V(G,p)_t^n = \frac{1}{\sqrt{n}\tau_n^p} \sum_{i=1}^{[nt]} \left(|\Delta_i^n G|^p - E|\Delta_i^n G|^p \right) ,$$

where $\Delta_i^n G = G_{\frac{i}{n}} - G_{\frac{i-1}{n}}$, $\tau_n^2 = E[|\Delta_i^n G|^2]$?

Problem 3: A more advanced problem is studied in [1]. Consider a process of the type

$$X_t = \int_{-\infty}^{t} g(t-s)\sigma_s dW_s ,$$

where $g \in L^2((0,\infty))$ is a memory function and σ is a stochastic process. Typically $g(x) \sim x^\alpha$ with $\alpha \in (-\frac{1}{2}, \frac{1}{2})$ (at 0), and thus X is not a semimartingale. The main interest is the derivation of the asymptotic behaviour (including CLT) of

$$V(X,p)_t^n = \frac{1}{\sqrt{n}\tau_n^p} \sum_{i=1}^{[nt]} |\Delta_i^n X|^p ,$$

where $\tau_n^2 = E[|\Delta_i^n G|^2]$ and $G_t = \int_{-\infty}^{t} g(t-s)dW_s$.

While *Problem 1* can be solved by classical methods (*diagram formula*), see e.g. [2], *Problem 2* and *Problem 3* are more complicated and require a different technique.

Recently, [3] and [4] proposed some new methods to prove CLT's on Gaussian fields that rely on operators of Malliavin calculus. Recall that any square integrable variable F on a Gaussian space has a unique Wiener chaos decomposition

$$F = \sum_{m=0}^{\infty} I_m(f_m) ,$$

where I_m is a multiple integral of order m and $f_m \in \mathbb{H}^{\odot m}$ are symmetric kernels. First of all, we need to understand how to prove a CLT for a *fixed* chaos.

Theorem ([3],[4]): Let $F_n = I_m(f_m^n)$ $(m \geq 1)$ be a sequence of random variables with

$$E[F_n^2] \to 1.$$

Let D denote the Malliavin derivative and let $g \otimes_p h$ be the pth contraction of $g, h \in \mathbb{H}^{\otimes m}$. The following conditions are equivalent:

(i) $F_n \Rightarrow N(0,1)$.

(ii) $E[F_n^4] \to 3$.

(iii) For all $1 \le p \le m-1$, $||f_m^n \otimes_p f_m^n||^2_{\mathbb{H}^{\otimes 2(m-p)}} \to 0$.

(iv) $||DF_n||^2_{\mathbb{H}} \to m$ in L^2.

This result can be extended to general sequences of the form $F_n = \sum_{m=d}^{\infty} I_m(f_m^n)$ ($d \ge 1$). In particular, the condition (iii) gives an easy method for proving CLT's in a rather general framework.

References

[1] O.E. Barndorff-Nielsen, J.M. Corcuera and M. Podolskij *Multipower variation for Brownian semistationary processes*, (2009). Working paper.

[2] P. Breuer, P. Major *Central limit theorems for nonlinear functionals of Gaussian fields*, Journal of Multivariate Analysis **13**(3) (1983), 425–441.

[3] D. Nualart, S. Ortiz-Latorre *Central limit theorems for multiple stochastic integrals and Malliavin calculus*, Stochastic Processes and Their Applications **118** (2008), 614–628.

[4] D. Nualart, G. Peccati *Central limit theorems for sequences of multiple stochastic integrals*, Annals of Probability **33** (2005), 177–193.

Pointwise adaptive estimation for robust and quantile regression

Markus Reiss

(joint work with Yves Rozenholc, Charles-André Cuenod)

We consider a generalized regression model

$$Y_i = g(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

with (ε_i) i.i.d., $x_1, \dots, x_n$ in the design space $\mathcal{X}$ and $g : \mathcal{X} \to R$. The problems we have in view are those of robust nonparametric estimation of g in the presence of heavy-tailed noise (ε_i) and of nonparametric quantile estimation, which is becoming more and more popular in applications. One main application is robust image denoising, in particular for sequences of CT images. In the spirit of classical M-estimation we consider $g(x_i)$ as the location parameter in the observation Y_i, that is

$$g(x_i) = \operatorname{arginf}_{m \in R} E[\rho(Y_i - m)] \tag{1}$$

for some convex function $\rho : R \to R^+$ with $\rho(0) = 0$. Standard examples are $\rho(x) = x^2/2$ for the classical mean regression model, $\rho(x) = |x|$ for the median regression model and more generally $\rho(x) = |x| + (2\alpha - 1)x$ for the quantile regression model. The function g is not supposed to satisfy a global smoothness criterion, but we aim at estimating it locally in each point $x \in \mathcal{X}$ as efficiently as possible. The risk will then depend on local regularity properties, which we do not assume to be known. For spatially inhomogeneous functions, in the presence of jumps or for image denoising pointwise adaptive methods are much more appropriate than global smoothing methods. In classical mean regression local adaptivity can be achieved using wavelet thresholding or kernels with locally varying bandwidths.

In this ideal situation a data-driven choice among linear empirical quantities is performed. M-estimators are typically nonlinear and the standard approaches do not necessarily transfer directly. Here, we develop a generic algorithm to select optimally among local M-estimators. In contrast to classical model selection, we do not only rely on the estimator values themselves to define a data-driven selection criterion. This has significant advantages in the present case of nonlinear base estimators.

Using Lepski's approach as a starting point, we argue in a multiple testing interpretation that our procedure is usually more powerful. Moreover, it is equally simple to analyze and easy to implement. We derive exact and asymptotic error bounds and the latter give optimal minimax rates for Hölder classes. Simulations show convincing finite sample properties. Moreover, they confirm that Lepski's classical method applied to local median estimators suffers from oversmoothing because changes in the signal are not detected early enough due to the robustness of the median. Finally, a dynamic extension of the procedure shows good results in denoising DCE-CT image sequences from cancer surveillance.

References

[1] M. Reiß, Y. Rozenholc, C.-A. Cuenod *Pointwise adaptive estimation for robust and quantile regression*, Preprint, Math arXive:0904.0543, 2009.

Data analysis of and on Dendrite Structures

Naoki Saito

Introduction: In this talk, we reported our preliminary results on a potential method to characterize dendrite structures of Retinal Ganglion Cells (RGCs) of mice by systematically extracting their morphological features based on their graph Laplacian eigenvalues so that we could save human interaction cost usually required for such feature extraction.

Analysis *of* Dendrite Structures: The segmentation and tracing software system used by our neuroscience collaborators provided us with a sequence of 3D coordinates that represent points sampled along dendrite arbors of RGCs with the branching information [2]. One of the most natural and simplest ways to model such a network-like structure is to construct a *graph*. Hence, our first task is to convert such a sequence of 3D points to a connected graph. Let G be a graph representing dendrite patterns of an RGC, let $V = V(G) = \{v_1, \ldots, v_n\}$ be a vertex set of G where each $v_k \in \mathbb{R}^3$ is a 3D sample point along dendrite arbors of this RGC, and let $E = E(G) = \{e_1, \ldots, e_m\}$ be an edge set of G where e_k connects two vertices v_i, v_j for some $1 \le i, j \le n$. Let d_{v_k} be the degree of the vertex v_k. In fact, dendrite pattern of each RGC in our dataset can be converted to a *tree* rather than a general graph since it is connected and contains no cycles. We also note that we only deal with *un*weighted graphs here. In other words, we examine the

connectivities and complexity of the dendrite graphs ignoring the physical lengths of the dendrite arbors.
Once we construct a graph per RGC, we proceed as follows:

Step 1: Construct the *Laplacian matrix* $L(G) := D(G) - A(G)$ where $D(G) := \mathrm{diag}(d_{v_1}, \ldots, d_{v_n})$ and $A(G) = (a_{i,j})$ is the adjacency matrix of G, i.e., $a_{i,j} = 1$ if v_i and v_j are adjacent; otherwise it is 0.
Step 2: Compute the eigenvalues of $L(G)$;
Step 3: Construct features using these eigenvalues;
Step 4: Repeat the above steps for all the RGCs and feed these feature vectors to a clustering algorithm of one's choice.

The Laplacian eigenvalues reflect various *intrinsic* geometric information about the graph e.g., connectivity, mean distance, etc.; see, e.g., [1, 3] for the details.
Let $|V| = n$, and let $0 = \lambda_0 \leq \lambda_1 \cdots \leq \lambda_{n-1}$ be the sorted eigenvalues of $L(G)$. Let $m_G(I)$ be the number of eigenvalues of $L(G)$, multiplicities included, that belong to a set $I \subset \mathbb{R}$. A vertex of degree 1 is called a *pendant* vertex, and a vertex adjacent to a pendant vertex is called *pendant neighbor*. Let $p(G)$ and $q(G)$ be the number of pendant vertices and the number of pendant neighbors of G, respectively. Let $i(G)$ be the *isoperimetric number* of G, which is closely related to the speed of convergence of a random walk on G to a stationary distribution. The *Wiener index* $W(G)$ of a graph G is the sum of the entries in the upper triangular part of the distance matrix $\Delta(G)$ of G, where $(\Delta(G))_{i,j}$ is the number of edges in a shortest path from v_i to v_j. The bounds of $q(G)$ and $i(G)$ as well as the exact value of $W(G)$ can be computed using the eigenvalues of $L(G)$; see [5].
We now report our preliminary results we obtained recently. The following features were used to characterize the dendrite patterns of 130 monostratified RGCs.

Feature 1: $(p(G) - m_G(1))/|V(G)|$ as a lower bound of the number of the pendant neighbors $q(G)$ with the normalization by $|V(G)|$;
Feature 2: The normalized Wiener index $W(G)/|V(G)|$;
Feature 3: $m_G(4, \infty)/|V(G)|$, i.e., the number of eigenvalues of $L(G)$ larger than 4 (normalized) ;
Feature 4: The upper bound of the isoperimetric number $i(G)$.

[5] explains why these features are used and shows the cross plots of these 4 features, from which we observe that sparsely and widely distributed dendrite patterns are well separated from the densely and narrowly distributed ones.

Discussion: We plan to analyze the Laplacian eigenvalues of the *weighted* graphs where the weight w_k of an edge $e_k = (v_i, v_j)$ is the inverse of the distance, i.e., $w_k = \|v_i - v_j\|^{-1}$ in our case, which should reflect more faithful geometric configuration of RGCs than those of the unweighted graphs. Analysis of such weighted graphs, however, are expected to be tougher because for example, $m_G(1)$ among the different RGCs does not have the same meaning anymore.
We also plan to analyze data distributed *on* such dendrite structures, e.g., temperature or density distributions, or ultimately, ionic current propagation. Here,

the *eigenfunctions* of the Laplacian will play a fundamental role since such distributions can be expanded into the Laplacian eigenfunctions, which permits one to do *spectral analysis and synthesis* of data on dendrite structures.
Finally, we will investigate *Poisson's equation with mixed boundary condition* on a dendrite pattern for characterizing the efficiency of information transmission of that neuron. It is well known (see, e.g., [4]) that the mean exit time $u(\boldsymbol{x})$ of particles released at a point $\boldsymbol{x}$ inside a bounded domain driven by Brownian motion is the solution of Poisson's equation $\Delta u = -1$ satisfying the zero Dirichlet boundary condition. We need to force the mixed boundary condition because the insulation along dendrites and axons leads to the Neumann boundary condition while the terminal regions (e.g., synapses and soma) lead to the Dirichlet boundary condition. Solving this Poisson's equation on such a domain itself is interesting, but it would be even more striking if we can extract features from its solution that are useful for charactering neurons.

References

[1] F. R. K. Chung, *Spectral Graph Theory*, no. 92 in CBMS Regional Conference Series in Mathematics, Amer. Math. Soc., Providence, RI, 1997.
[2] J. Coombs, D. van der List, G.-Y. Wang, and L. M. Chalupa, *Morphological properties of mouse retinal ganglion cells*, Neuroscience, 140 (2006), pp. 123–136.
[3] R. Merris, *Laplacian matrices of graphs: A survey*, Linear Algebra Appl., 197/198 (1994), pp. 143–176.
[4] M. A. Pinsky, *Mean exit time from a bumpy sphere*, Proc. Amer. Math. Soc., 122 (1994), pp. 881–883.
[5] N. Saito and E. Woei, *Analysis of neuronal dendrite patterns using eigenvalues of graph Laplacians*, JSIAM Letters, 1 (2009), pp. 13–16. Invited paper.

Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density

Richard Samworth
(joint work with Madeleine Cule)

Shape-constrained density estimation in general, and log-concave density estimation in particular, have received a great deal of attention in the statistical literature recently – see, for example, [7], [3], [4], [6], [5], [1]. The following theorem given in [3] helps to explain this interest:

Theorem 4. *Let $X_1, \ldots, X_n$ be independent with density f_0 in $\mathbb{R}^d$, and suppose that $n \geq d+1$. Then, with probability one, there exists a unique log-concave maximum likelihood estimator $\hat{f}_n$ of f_0.*

Thus, even though the class of log-concave densities is infinite-dimensional (and contains many well-known and commonly-used families of densities), there exists a fully automatic density estimator within this class, with no smoothing parameters to choose. To understand the theoretical properties of this estimator, we begin by noting that when it is known that a sequence of densities is log-concave,

convergence in weak sesnses in fact implies convergence in much stronger senses (see [2]):

Proposition 5. *Let (f_n) be a sequence of log-concave densities on $\mathbb{R}^d$ with $f_n \xrightarrow{d} f$ for some density f. Then:*

(a) *f is log-concave*
(b) *$f_n \to f$, almost everywhere*
(c) *Let $a_0 > 0$ and $b_0 \in \mathbb{R}$ be such that $f(x) \leq e^{-a_0\|x\|+b_0}$. Then for every $a < a_0$, we have $\int_{\mathbb{R}^d} e^{a\|x\|}|f_n(x) - f(x)|\,dx \to 0$ and, if f is continuous, $\sup_{x\in\mathbb{R}^d} e^{a\|x\|}|f_n(x) - f(x)| \to 0$.*

In order to state our main result, we first require appropriate bounds on the behaviour of the log-concave maximum likelihood estimator, as illustrated in the following result that can be found in [2]. Write E for the support of f_0.

Lemma 6. *Suppose that $\int_{\mathbb{R}^d} \|x\| f_0(x)\,dx < \infty$.*

(a) *There exists a constant $C > 0$ such that, with probability one,*

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \hat{f}_n(x) \leq C.$$

(b) *Let S be a compact subset of $\mathrm{int}(E)$. There exists a constant $c > 0$ such that, with probability one,*

$$\liminf_{n\to\infty} \inf_{x\in\mathrm{conv}\,S} \hat{f}_n(x) \geq c.$$

Our main result establishes establishes desirable performance properties of $\hat{f}_n$. Recall that the Kullback–Leibler divergence of a density f from f_0 is given by $d_{KL}(f_0, f) = \int_{\mathbb{R}^d} f_0 \log(f_0/f)$. Jensen's inequality shows that the Kullback–Leibler divergence is non-negative, and equal to zero if and only if $f = f_0$ (almost everywhere). Thus when f_0 is log-concave, Theorem 7 below shows that the log-concave maximum likelihood estimator $\hat{f}_n$ is strongly consistent in certain exponentially weighted total variation metrics. Convergence in exponentially weighted supremum norms also follows if f_0 is continuous.

In the case where the model is misspecified, i.e. f_0 is not log-concave, we prove that the existence and uniqueness of a log-concave density f^* that minimises the Kullback–Leibler divergence from f_0. Moreover, we show that the log-concave maximum likelihood estimator $\hat{f}_n$ converges in the same senses as in the previous paragraph to f^* (see [2] for details). We write $\log_+ x = \max(\log x, 0)$.

Theorem 7. *Let f_0 be any density on $\mathbb{R}^d$ with*

$$\int_{\mathbb{R}^d} \|x\| f_0(x)\,dx < \infty, \quad \int_{\mathbb{R}^d} f_0 \log_+ f_0 < \infty \quad \text{and} \quad \mathrm{int}(E) \neq \emptyset.$$

There exists a log-concave density f^*, unique almost everywhere, that minimises the Kullback–Leibler divergence of f from f_0 over all log-concave densities f. Taking $a_0 > 0$ and $b_0 \in \mathbb{R}$ such that $f^*(x) \leq e^{-a_0\|x\|+b_0}$, we have for any $a < a_0$

that

$$\int_{\mathbb{R}^d} e^{a\|x\|} |\hat{f}_n(x) - f^*(x)|\, dx \overset{a.s.}{\to} 0,$$

and, if f^* *is continuous,* $\sup_{x\in\mathbb{R}^d} e^{a\|x\|} |\hat{f}_n(x) - f^*(x)| \overset{a.s.}{\to} 0$.

References

[1] Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009), Limit distribution theory for maximum likelihood estimation of a log-concave density, *Ann. Statist.*, **37**, 1299–1331.

[2] Cule, M. L. and Samworth, R. J. (2009), *Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density.* Preprint. Available at `http://arxiv.org/pdf/0908.4400`.

[3] Cule, M. L., Samworth, R. J. and Stewart, M. I. (2007), *Maximum likelihood estimation of a multidimensional log-concave density*, Submitted. Available at `http://www.statslab.cam.ac.uk/~rjs57/Research.html`.

[4] Dümbgen, L., Hüsler, A. and Rufibach, K. (2007), Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data. Preprint. `http://arxiv.org/abs/0707.4643`.

[5] Dümbgen, L. and Rufibach, K. (2009), Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency, *Bernoulli*, **15**, 40–68.

[6] Pal, J., Woodroofe, M. and Meyer, M. (2007), Estimating a Polya frequency function. In *Complex datasets and Inverse problems: Tomography, Networks and Beyond* (eds. R. Liu, W. Strawderman, C.-H. Zhang), pp. 239–249. IMS Lecture Notes - Monograph Series **54**.

[7] Walther, G. (2002), Detecting the presence of mixing with multiscale maximum likelihood, *J. Amer. Statist. Assoc.*, **97**, 508–513.

Estimation of convex-transformed densities

Arseni Seregin

(joint work with Jon A. Wellner)

Density estimation with shape or geometric constraints lies between the unconstrained (such as kernel or regularized) estimation and parametric estimation, and has advantages of both. It allows for very large classes of densities and it does not require a choice of a metaparameter (such as bandwidth in kernel estimation). In many cases the MLE estimator exists, is consistent and rate-optimal. Finally, such estimators tend to adapt to the smoothness of the true density.

Our work is an example of shape constrained density estimation where we exploit convexity as a geometric constraint. We study the properties of the nonparametric maximum likelihood estimator (MLE) of a convex-transformed density i.e. the density of the form $f \circ h_0$, where h_0 is an arbitrary convex function on $\mathbb{R}^d$ and $f : \mathbb{R} \to \mathbb{R}_+$ is a known monotone transformation which defines the class $\mathcal{M}(f)$ of such densities. We distinguish two types of such classes: decreasing and increasing models which correspond to decreasing and increasing transformations.

A decreasing model consists of unimodal densities with convex superlevel sets. Changing the transformation f we can obtain density functions with 'heavy' or 'light' tails. On the other hand, changing the convex function h_0 we can vary the geometry of superlevel sets. As an example we can consider the decreasing

transformation $f(y) = \exp(-y)$ which provides the well-known class of log-concave densities. The global rates and consistency of the MLE for log-concave model when $d = 1$ are studied in Dümbgen and Rufibach [2] and Balabdaoui, Rufibach and Wellner[1] establish the local rates of convergence. However, for the classes with heavier tails no results about the MLE behavior are known even in one-dimensional case. As an alternative approach, Koenker and Mizera in the unpublished paper [4] show the existence and Fisher consistency of regularized estimators for the models which correspond to our decreasing models with convex transformation f.

One can show that non-degenerate increasing model cannot contain densities defined on the whole space. In order to define a density in such a model we first need to define a convex domain of h_0. We restrict the domain of h to a positive orthant $\mathbb{R}^d_+$. As an example of an increasing model we can consider a one-dimensional model defined by the transformation $f(y) = \max(y, 0)$. When $d = 1$ this family is equivalent to the family of decreasing convex densities which was studied in Groeneboom, Jongbloed and Wellner [3].

Under mild conditions on tail behavior the transformation f we prove that the MLE exists almost surely when the number of observations is large enough for both decreasing and increasing models. We also prove that the MLE for the decreasing model is Hellinger and pointwise consistent. Finally, we establish the minimax lower bounds for the density estimation at a given point and for estimation of the mode.

References

[1] F. Balabdaoui, K. Rufibach and J. A. Wellner *Limit distribution theory for maximum likelihood estimation of a log-concave density*, Ann. Statist. **37** (2009), 1299-1331.

[2] L. Dümbgen and K. Rufibach *Maximum likelihood estimation of a log-concave density and its density and its distribution function*, Tech. rep., University of Bern (2007), Available at arXiv:0709.0334.

[3] P. Groeneboom, G. Jongbloed, and J. A. Wellner *Estimation of a convex function: characterizations and asymptotic theory*, Ann. Statist. **29** (2001), 1653-1698.

[4] R. Koenker, and I. Mizera *Quasi-concave density estimation*, unpublished (2008).

Smoothed absolute loadings principal components analysis

Bernard W. Silverman

A crucial part of genome-wide association studies is the identification of modes of variability in genome data which do not depend on small parts of the genome. The basic data we consider is an $n \times p$ data matrix X whose rows represent individuals and columns SNPs (single nucleotide polymorphisms). Each SNP can take values coded $0-1-2$, but for each SNP there is an underlying arbitariness in the direction of coding. Let Y be the matrix X with column means subtracted.

The natural statistical starting-point is principal components analysis, but in practice raw principal components produce loadings concentrated on a small number of SNPs, as shown in the Figure. Therefore some sort of regularization is required. Methods such as Silverman (1996) work by controlling the amount of

local variability in the loadings vector u, but this is not appropriate in the current case, because of the arbitrary coding of the individual SNPs. It only makes sense to to consider regularization approaches which depend on the absolute values of the loadings.

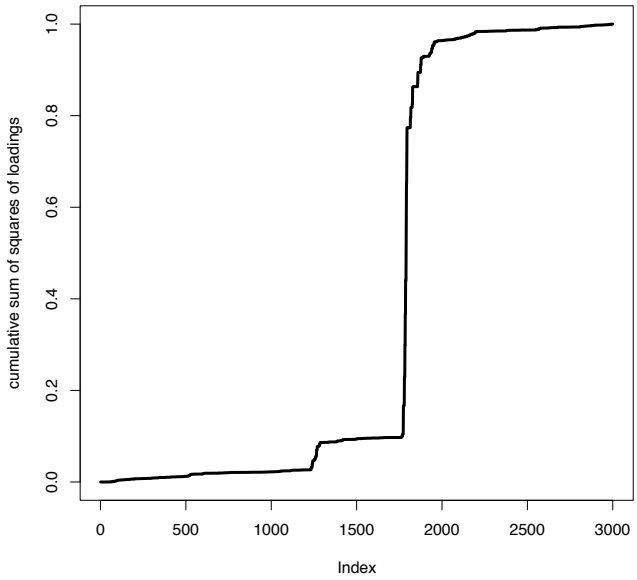

FIGURE 1. Cumulative sum of squares of loadings of the first principal component of a 3000 SNP part of Chromosome 22 on 1000 individuals

The simplest approach is to add to standard PCA the constraint that all the $|u_j|$ are equal. This is essentially equivalent to the *centroid* method, which replaces the usual constraint $\|u\|_2 = 1$ by $u_j = \pm 1$ for all j. The method dates back to Burt (1917) and Thurstone (1931), but has been the focus of recent attention by Choulakian and co-authors. For example, Choulakian (2006) shows that a local optimum of the centroid method is obtained by the simple recursion $u \leftarrow \text{sign } Y'Yu$ iterated to convergence from a suitable starting value.

This approach is an extreme form of smoothing of the absolute loadings. A more flexible methodology is obtained by enforcing some regularity on the $\|u_j\|$, but not forcing them all to be equal. Let $\mathcal{B}$ be a space of 'smooth' p-vectors b; the space being considered in current exploratory work is the vector of values at the integers $0, 1, \ldots, p-1$ of functions generated by a B-spline basis on $[0, p-1]$. Now consider the problem

$$\max \|Yu\|_2^2 / \|u\|_2^2 \text{ subject to } |u| \in \mathcal{B} \tag{1}$$

where $|u|$ is the vector whose elements are the absolute values $|u_i|$. Optimization problem (1) adds to standard PCA the additional requirement that the vector of *absolute* loadings falls in the space $\mathcal{B}$ of 'smooth' vectors.

We approach (1) by an alternating maximization approach. For any given vector of absolute values of the loadings, an adaptation of the centroid method aims to choose the signs of the loadings optimally. On the other hand, for suitable spaces $\mathcal{B}$

we can find the best loadings for any given sequence $\epsilon = \text{sign } u$. Each of these two steps—finding the best signs for given absolute values, or finding the best absolute values for given signs—increases the value of the objective function $\|Yu\|_2^2/\|u\|_2^2$ and hence iterating the alternate maximization leads to a (local) maximum of the problem (1).

There are many interesting questions raised by this approach. From a theoretical point of view, it would be instructive to investigate the asymptotic properties of the approach under suitable assumptions, under which all three of n and p and the order of the B-splines would vary in the limiting regime considered, though even the centroid method itself does not appear to have been studied in this way. There are obvious methodological questions; for example, is there a reasonable cross-validation approach to the choice of the family $\mathcal{B}$ via the parameters (degree and number of knots) in the B-spline basis? Other questions are computational, for example the appropriate choice of starting vector for the alternating maximisation algorithm. One fascinating possibility is to use a genetic algorithm where new starting points are obtained by appropriate 'breeding' of current good estimates of the loading vector. Preliminary experiments suggest this approach is promising. Another computational challenge is to cast the problem is a form amenable for parallel computation and very large data sets. In many current applications we may be looking at millions of SNPs and thousands of individuals, so the data matrix X is very large indeed, and into the future the size of appropriate genetic data sets will certainly increase.

Finally, of course, will be the question of whether this approach to extracting modes of variability in genetic data sets yields a useful input into the pressing problems for which the data were originally collected.

References

[1] C. Burt, The Distribution and Relations of Educational Abilities. London: P. S. King & Son. (1917)

[2] V. Choulakian, *L_1-norm projection pursuit principal component analysis*, Computational Statistics and Data Analysis **50** (2006), 1441–1451.

[3] B. W. Silverman, *Smoothed functional principal components analysis by choice of norm*, Annals of Statistics **24** (1996), 1–24.

[4] L. L. Thurstone, *Multiple Factor Analysis*, Psychological Review **38** (1931), 406–427.

Statistical inference for stochastic coefficient regression models

Suhasini Subba Rao

The classical multiple linear regression model is ubiquitous in many fields of research. However, in situations where the response variable $\{Y_t\}$ is observed over time, it is not always possible to assume that the influence the regressors $\{x_{t,j}\}$ exert on the response Y_t is constant over time. In order to allow for the influence of the previous regression coefficient on the current coefficient it is often reasonable to assume that the underlying unobservable regression coefficients are stationary processes and each coefficient admits a linear process representation. In other

words, a plausible model for modelling the varying influence of regressors on the response variable is

$$Y_t = \sum_{j=1}^{n} (a_{j,0} + \alpha_{t,j}) x_{t,j} + \varepsilon_t = \sum_{j=1}^{n} a_{j,0} x_{t,j} + X_t,$$

where $\{x_{t,j}\}$ are the deterministic regressors, $\{a_{j,0}\}$ are the mean regressor coefficients, $\mathrm{E}(X_t) = 0$ and satisfies $X_t = \sum_{j=1}^{n} \alpha_{t,j} x_{t,j} + \varepsilon_t$, $\{\varepsilon_t\}$ and $\{\alpha_{t,j}\}$ are jointly stationary linear time series with $\mathrm{E}(\alpha_{t,j}) = 0$, $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{E}(\alpha_{t,j}^2) < \infty$ and $\mathrm{E}(\varepsilon_t^2) < \infty$. We observe that this model includes the classical multiple regression model as a special case, with $\mathrm{E}(\alpha_{t,j}) = 0$ and $\mathrm{var}(\alpha_{t,j}) = 0$. The above model is often refered to as a stochastic coefficient regression (SCR) model. Such models have a long history in statistics (see, for example, Hilderth and Houck (1968) Burnett and Gutherie (1970), Newbold and Bos (1985) and Franke and Gründer (1995)).

Before fitting an SCR model to the data, it is of interest to check whether there is any evidence to suggest the coefficients are random and correlated. Let us consider the null hypothesis of fixed coefficients $H_0 : Y_t = a_0 + a_1 x_t + \epsilon_t$ where $\{\epsilon_t\}$ are iid random variables with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{var}(\varepsilon_t) = \sigma_\epsilon^2 < \infty$ against the alternative of random coefficients $H_A : Y_t = a_0 + a_1 x_t + \epsilon_t$, where $\epsilon_t = \alpha_t x_t + \varepsilon_t$ and $\{\alpha_t\}$ and $\{\varepsilon_t\}$ are iid random variables with $\mathrm{E}(\alpha_t) = 0$, $\mathrm{E}(\varepsilon_t) = 0$, $\mathrm{var}(\alpha_t) = \sigma_\alpha^2 < \infty$ and $\mathrm{var}(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$. We observe if the alternative were true, then $\mathrm{var}(\epsilon_t) = x_t^2 \sigma_\alpha^2 + \sigma_\varepsilon^2$, hence plotting $\mathrm{var}(\epsilon_t)$ against x_t should give a clear positive slope. Using this observation, one can construct a test for fixed regression parameters using a test statistic which is the sample correlation between $\{x_t^2\}$ and the empirical residuals $\{\hat{\epsilon}_t^2\}$. A similar test can be constructed to test for randomness of the coefficients.

We now consider methods for estimating the parameters in the SCR model. In the aforementioned literature, it is usually assumed that $\{\alpha_{t,j}\}$ satisfies a parametric linear time series model and the Gaussian maximum likelihood (GML) is used to estimate the unknown parameters. In the case $\{Y_t\}$ is Gaussian, the estimators are asymptotically normal and the variance of these estimators can be obtained from the inverse of the Information matrix. Even in the situation $\{Y_t\}$ is non-Gaussian, the Gaussian likelihood is usually used as the objective function to be maximised, in this case the objective function is often called the quasi-Gaussian likelhood (quasi-GML). The quasi-GML estimator is a consistent estimate of the parameters (see Caines (1988), Chapter 8.6) but when $\{Y_t\}$ is non-Gaussian, obtaining an expression for the standard errors of the quasi-GML estimators seems to be almost imposible. Therefore implicitly it is usually assumed that $\{Y_t\}$ is Gaussian, and most statistical inference is based on the assumption of Gaussianity. In several situations the assumption of Gaussianity may not be plausible, and there is a need for estimators which are free of distributional assumptions. To address this issue we propose an alternative, in some sense nonparametric estimator. Let $\boldsymbol{a}$ denote the mean regression parameters and $\boldsymbol{\theta}$ the parameters which characterise

the impulse response sequences of the linear processes of the stochastic coefficients and error. We use $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T) = \arg\min \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta})$ as an estimator of $(\boldsymbol{a}, \boldsymbol{\theta})$, where

$$(1) \quad \mathcal{L}_T^{(m)}(\boldsymbol{a}, \boldsymbol{\theta}) = \frac{1}{T-m} \sum_{t=m/2}^{T-m/2} \int_{-\pi}^{\pi} \left\{ \frac{\mathcal{I}_{t,m}(\boldsymbol{a}, \omega)}{\mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega)} + \log \mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega) \right\} d\omega,$$

m is even,

$$\mathcal{I}_{t,m}(\boldsymbol{a}, \omega) = \frac{1}{2\pi m} \Big| \sum_{k=1}^{m} (Y_{t-m/2+k} - \sum_{j=1}^{n} a_j x_{t-m/2+k,j}) \exp(ik\omega) \Big|^2$$

and

$$\mathcal{F}_{t,m}(\boldsymbol{\theta}, \omega) = \sum_{j=1}^{n} \sigma_j^2 \int_{-\pi}^{\pi} I_{t,m}^{(j)}(\lambda) f_j(\boldsymbol{\theta}, \omega - \lambda) d\lambda + \sigma_{n+1}^2 \int_{-\pi}^{\pi} I_m^{(n+1)}(\lambda) f_{n+1}(\boldsymbol{\theta}, \omega - \lambda) d\lambda$$

We also consider a closely related estimator. Both of the proposed methods offer an alternative perspective of the SCR model based within the frequency domain, and are free of any distributional assumptions. The asymptotic sampling properties such as consistency and asymptotic normality can be derived. In particular, the variance of the asymptotic distribution of $(\hat{\boldsymbol{a}}_T, \hat{\boldsymbol{\theta}}_T)$ can be derived when the distribution of the stochastic coefficients and errors are unknown. A theoretical comparison of our estimators with the GML estimator, in most cases, is not possible, this is because it is usually not possible to obtain the asymptotic variance of the GML estimator. However, in the case that the random coefficients and errors are Gaussian it is possible to show that GML estimator and our frequency domain estimator have asymptotically equivalent distributions. Details can be found in Subba Rao (2009).

References

[1] T. D. Burnett, T. D. and D. Guthrie, *Estimation of stationary stochastic regression parameters*, J. American Statistical Association **32** (1990), 120–140.

[2] P. Caines, Linear stochastic systems, Wiley, 1988.

[3] J. Franke, and B. Gründer, *General kriging for spatial-temporal processes with random ARX-regression parameters*, Athens conference in Applied Probability and Time Series Analysis: Vol II, 1995, 177-189.

[4] C. Hildreth and C. Houck, *Some estimates for a linear model with random coefficients*, J. of the American Statistical Association **63** (1968), 584–595.

[5] P. Newbold, and T. Bos, Stochastic parameter regression models, Sage Publications.

[6] D. S. Stoffer, and K. D. Wall, *Bootstrappping state space models: Gaussian maximum likelihood estimation*, J. of the American Statistical Association **86** (1991), 1024-1033.

[7] S. Subba Rao, *Statistical inference for stochastic coefficient regression models*, preprint (2009).

Estimation of Jump Tails

Viktor Todorov
(joint work with Tim Bollerslev)

We consider the estimation of the jump tails of a financial asset price, whose logarithm is assumed to be an Ito semimartingale with the following dynamics

$$dp_t = \alpha_t dt + \sigma_t dW_t + \int_{\mathbb{R}} \kappa(x)\tilde{\mu}(dt,dx) + \int_{\mathbb{R}} \kappa'(x)\mu(dt,dx), \tag{1}$$

where α_t, σ_t are some locally bounded processes and W_t is a Brownian motion; μ is a one-dimensional measure on $[0,\infty)\times\mathbb{R}$ counting the number of jumps of given size over a given interval of time; the compensator of the jump measure is denoted with $\nu_t(dx)dt$ and $\tilde{\mu}(dt,dx) := \mu(dt,dx) - \nu_t(dx)dt$ is the compensated measure; $\kappa(x)$ is a bounded continuous function which equals x around the origin and $\kappa'(x) = x - \kappa(x)$.

We assume that the compensator of the jumps $\nu_t(x)$ has the following decomposition

$$\nu_t(x) = \left[(k_0^+ + k_1^+\sigma_t^2)1_{x>0} + (k_0^- + k_1^-\sigma_t^2)1_{x<0}\right]\nu(x), \tag{2}$$

where $\nu(x)$ is a Levy measure with regularly varying tails, and some further regularity assumptions being fulfilled.

On a first step of the analysis we show how to conduct the estimation when we have a continuous record of the price. The idea of the estimation can be described as follows. We first construct moment conditions based on the unobserved compensated measure ν_t. For this we use the scores of a generalized Pareto distribution for the excesses of jumps over a given threshold (which in general goes to zero asymptotically). The latter are valid scores asymptotically because of our assumption of regularly varying tails (see e.g. [3]). To make the moment conditions feasible, next we replace the measure ν_t in the moment conditions with the measure μ, which we observe: the difference is a martingale that does not affect the moments. Finally, to identify the constant and time varying part of ν_t we use projection on lagged $\int_t^{t+1}\sigma_s^2 ds$.

The second part of the analysis consists of extending the above estimation results to the situation when the data is observed discretely over a grid whose mesh goes to zero. In particular, we assume that we observe the log-price p_t at times $0, \Delta_n, ..., [T/\Delta_n]\Delta_n$, where T is the span of the data set and Δ_n is the length of the high-frequency intervals. To adapt the "infeasible" estimation to the current discrete setting, we develop estimators from the high-frequency data of integrals with respect to the jump measure as well as of the integrated volatility.

Our estimator of integrals of the form $\int_t^{t+1}\int_{\mathbb{R}}\phi(x)\mu(ds,dx)$ is:

$$\sum_{i=1}^{[1/\Delta_n]} \phi(\Delta_i^{n,t}p)1_{\{|\Delta_i^{n,t}p|\geq\alpha\Delta_n^{\varpi}\}}, \tag{3}$$

where $\Delta_i^{n,t} p := p_{t+i\Delta_n} - p_{t+(i-1)\Delta_n}$, $\alpha > 0$, $\varpi \in (0, \frac{1}{2})$. Similarly, our estimate of $\int_t^{t+1} \sigma_s^2 ds$ is the truncated variance of [2] (see also [1]):

$$\sum_{i=1}^{[1/\Delta_n]} \left(\Delta_i^{n,t} p\right)^2 1_{\{|\Delta_i^{n,t} p| \le \alpha \Delta_n^{\varpi}\}}. \tag{4}$$

We substitute these estimates from the high-frequency data in the estimation equations derived for the continuous-record sampling scheme. Then we show that under certain conditions for the relative speed of $\Delta_n \downarrow 0$ when compared with $T \uparrow \infty$, we have that the estimator based on the high-frequency data is asymptotically equivalent to the one based on the continuous record. These conditions depend on: (1) the maximum power p for which $\mathbb{E}|\sigma_t|^p$, (2) the Blumenthal-Getoor index of the jumps, and (3) how fat are the jump tails.

We conclude with showing how and when our estimation results can be extended to the case when the jump compensator $\nu_t(x)$ is given by the more general specification

$$\nu_t(dx) = \left(\nu_0(x) + \nu_1(x)\sigma_t^2\right) dx, \quad \text{for every } t \text{ and } x \text{ big enough}, \tag{5}$$

where ν_0 and ν_1 are two valid Levy measures with regularly varying tails. The conditions limit the possible difference in the tail behavior of ν_0 and ν_1. These conditions are stronger, the bigger is the deviation of the tails from power law.

References

[1] J. Jacod, *Asymptotic Properties of Power Variations and Associated Functionals of Semimartingales*, Stochastic Processes and their Applications **118** (2008), 517–559.

[2] C. Mancini, *Nonparametric Threshold Estimation for Models with Stochastic Diffusion Coefficient and Jumps*, Scandinavian Journal of Statistics **36** (2009), 270–296.

[3] R. Smith, *Estimating Tails of Probability Distributions*, Annals of Statistics **15** (1987), 1174–1207.

Feature Variables in High-Dimensional Linear Regression Stepwise Searching in High-Dimensional Regression

Qiwei Yao

(joint work with Cun-Hui Zhang, Da Huang and Hongzhi An)

We investigate the classical stepwise forward and backward search methods for selecting sparse models in the context of linear regression with the number of candidate variables p greater than the number of observations n. Two types of new information criteria BICP and BICC are proposed to serve as the stopping rules in the stepwise searches, since the traditional information criteria such as BIC and AIC are designed for the cases with $p < n$, and may fail spectacularly when p is close to or greater than n. The proposed methods are illustrated in a simulation study which indicates that the new methods outperform a counterpart LASSO selector with a penalty parameter set at a fixed value. The consistency of the stepwise search is investigated when both n and p tend to ∞. We show

that a stepwise forward addition followed by a stepwise backward deletion, both controlled by a version of BICP, leads to a consistent estimated model under the sparse Riesz condition.

Reporter: Wolfgang Polonik

Participants

Prof. Dr. Ethan Anderes
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Dragi Anevski
Dept. of Mathematics
University of Lund
Box 118
S-221 00 Lund

Prof. Dr. Rudolf J. Beran
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Lucien Birge
Laboratoire de Probabilites-Tour 56
Universite P. et M. Curie
4, Place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
CH-8092 Zürich

Prof. Dr. Rui Castro
Columbia University
Electrical Engineering Department
500 West 120th Street
New York , NY 10027
USA

Prof. Dr. Claudia Czado
Lehrstuhl für Mathematische
Statistik
Technische Universität München
85747 Garching bei München

Prof. Dr. Richard A. Davis
Department of Statistics
Columbia University
1255 Amsterdam Avenue, MC 4690
Room 1010 SSW
New York , NY 10027
USA

Dr. Helmut Finner
DDZ
Institut für Biometrie und Epidemiologie
Auf'm Hennekamp 65
40225 Düsseldorf

Prof. Dr. Marc Genton
Department of Statistics
Texas A & M University
College Station , TX 77843-3143
USA

Dr. Sonja Greven
Department of Biostatistics
School of Hygiene and Public Health
Johns Hopkins University
615 North Wolfe Street
Baltimore , MD 21205
USA

Prof. Dr. Torsten Hothorn
Institut für Statistik
Universität München
Ludwigstr. 33
80539 München

Prof. Dr. Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology
765 Ferst Drive
Atlanta GA 30332-0205
USA

Prof. Dr. Marie Huskova
Department of Probability and Mathematical Statistics
Charles University
Sokolovska 83
18675 Praha 8
CZECH REPUBLIC

Prof. Dr. Thomas Klein
Zentrum Mathematik
TU München
Boltzmannstr. 3
85748 Garching bei München

Prof. Dr. Claudia Klüppelberg
Zentrum Mathematik
TU München
Boltzmannstr. 3
85748 Garching bei München

Prof. Dr. Samuel Kou
Department of Statistics
Harvard University
One Oxford Street
Cambridge , MA 02138
USA

Eric Laber
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
USA

Prof. Dr. Alexander Lindner
Institut für Mathematische Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig

Prof. Dr. Yanyuan Ma
Department of Statistics
Texas A & M University
College Station , TX 77843-3143
USA

Prof. Dr. David M. Mason
Department of Food and Resource Economics
University of Delaware
206 Townsend Hall
Newark DE 19717
USA

Dr. Nicolai Meinshausen
Department of Statistics
University of Oxford
1 South Parks Road
GB-Oxford OX1 3TG

Prof. Dr. Thomas Mikosch
Laboratory of Actuarial Mathematics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen

Dr. Aleksey Min
Zentrum Mathematik
TU München
Boltzmannstr. 3
85748 Garching bei München

Prof. Dr. Ivan Mizera
Dept. of Mathematics and Statistics
University of Alberta
632 Central Academic Bldg.
Edmonton, AB T6G 2G1
CANADA

Dr. Gernot Müller
Zentrum Mathematik
TU München
Boltzmannstr. 3
85748 Garching bei München

Prof. Dr. Klaus-Robert Müller
Technical University of Berlin
Department Computer Science
Franklinstr. 28/29
10587 Berlin

Prof. Dr. Axel Munk
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Prof. Dr. Susan A. Murphy
Department of Statistics
University of Michigan
439 West Hall
1085 South University
Ann Arbor MI 48109-1107
USA

Prof. Dr. Robert Nowak
University of Wisconsin-Madison
3627 Engineering Hall
1415 Engineering Drive
Madison , WI 53706
USA

Prof. Dr. Liam Paninski
Dept. of Statistics
Columbia University
New York , NY 10027
USA

Dr. Mark Podolskij
Departement Mathematik
ETH-Zentrum
Rämistr. 101
CH-8092 Zürich

Prof. Dr. Wolfgang Polonik
Department of Statistics
University of California, Davis
One Shields Avenue
Davis CA 95616
USA

Prof. Dr. Suhasini Subba Rao
Department of Statistics
Texas A & M University
College Station , TX 77843-3143
USA

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

Prof. Dr. Yaacov Ritov
Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905
ISRAEL

Prof. Dr. Naoki Saito
Department of Mathematics
University of California, Davis
1, Shields Avenue
Davis , CA 95616-8633
USA

Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
GB-Cambridge CB3 OWB

Dr. Günther Sawitzki
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg

Jürg Schelldorfer
Departement Mathematik
ETH-Zentrum
Rämistr. 101
CH-8092 Zürich

Johannes Schmidt-Hieber
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen

Arseni Seregin
Department of Statistics
University of Washington
Box 35 43 22
Seattle , WA 98195-4322
USA

Prof. Dr. Bernard W. Silverman
St. Peter's College
University of Oxford
GB-Oxford OX1 2DL

Prof. Dr. Vladimir Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin

Dr. Robert Stelzer
Zentrum für Mathematik
Technische Universität München
Boltzmannstr. 3
85748 Garching bei München

Prof. Dr. Viktor Todorov
Kellogg School of Management
Northwestern University
Department of Finance
Evanston IL 60208
USA

Prof. Dr. Alexandre B. Tsybakov
Laboratoire de Probabilites
Universite Paris 6
4 place Jussieu
F-75252 Paris Cedex 05

Prof. Dr. Jon A. Wellner
Department of Statistics
University of Washington
Box 35 43 22
Seattle , WA 98195-4322
USA

Prof. Dr. Qiwei Yao
Department of Statistics
London School of Economics
Houghton Street
GB-London WC2A 2AE