Mathematisches Forschungsinstitut Oberwolfach

Report No. 48/2011

DOI: 10.4171/OWR/2011/48

Very High Dimensional Semiparametric Models

Organised by
Arnold Janssen, Düsseldorf
Aad W. van der Vaart, Amsterdam
Jon A. Wellner, Seattle

October 2nd – October 8th, 2011

Abstract. Very high dimensional semiparametric models play a major role in many areas, in particular in signal detection problems when sparse signals or sparse events are hidden among high dimensional noise. Concrete examples are genomic studies in biostatistics or imaging problems. In a broad context all kind of statistical inference and model selection problems were discussed for high dimensional data.

Mathematics Subject Classification (2000): 62G10, 62G05, 62J15, 62F15, 62G20.

Introduction by the Organisers

The workshop *Very High Dimensional Semiparametric Models*, organised by Arnold Janssen (Düsseldorf), Aad W. van der Vaart (Amsterdam) and Jon A. Wellner (Seattle) was held October 2nd– October 8th, 2011. It was well attended with 52 participants from 11 countries from different continents. This workshop was a nice blend of researchers with various statistical backgrounds.
The talks covered a broad spectrum from modern statistical theory for very high dimensional problems. During the week 27 talks were given including 5 extended morning talks about outstanding topics. Throughout much time was spent for long and lively discussions. Special topics were:

- The sparsity in high dimensions with applications in medicine, biology and astronomy.

- Bayesian methods and reduction of dimension including regularisation methods, computation and penalty functions for estimation problems.
- Qualitative assumptions about monotonicity and convexity.
- Beyond the parametric boundary topics about estimation and the bias-variance trade-off.

It was also very successful to bring more applied researchers together with colleagues from mathematical statistics. This combination was very stimulating for further research and discussion. In particular a lot of young researchers attended the conference. The meeting was a great success. As always the stimulating atmosphere of the Forschungsinstitut led to an extensive exchange of ideas. A lot of new scientific contacts were formed, initiating quite a number of collaborations.

Workshop: Very High Dimensional Semiparametric Models

Table of Contents

Abstracts

Estimating a composite function by Model Selection

Yannick Baraud
(joint work with Lucien Birgé)

We consider an n-sample $X_1, \dots, X_n$ with values in $[-1,1]^k$ of common density s^2. Our aim is to estimate the function s with the L_2-loss when k is large and, to do so, look for some best approximation by composite functions of the form $g \circ u$. Our solution is based on model selection and leads to a very general approach to solve this problem with respect too many different types of functions g, u. In particular, we handle the problems of approximating s by additive functions, single and multiple index models. We also investigate the situation where $s = g \circ u$ for functions g and u belonging to possibly anisotropic smoothness classes. In this case, our approach leads to a completely adaptive estimator with respect to the regularity of s.

References

[1] J. Horowitz; E. Mammen, *Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions.*, Ann. Statist. **35** (2007).

[2] A. Juditsky; O. Lepski; A. Tsybakov *Nonparametric estimation of composite functions*, Ann. Statist. **37** (2009).

High-dimensional causal inference

Peter Bühlmann

We discuss causal inference when the number of variables may be much larger than sample size. Sparsity of the underlying directed acyclic graph is crucial for estimation accuracy and improved identifiability. We illustrate potential and limitations of high-dimensional causal inference, and we show some applications in genomics.

Minimax and Adaptive Estimation of Large Covariance Matrices

Tony Cai

Covariance structure is of fundamental importance in many areas of statistical inference and a wide range of applications, including genomics, fMRI analysis, risk management, and web search problems. In the high dimensional setting where the dimension p can be much larger than the sample size n, classical methods and results based on fixed p and large n are no longer applicable. In this talk, I discuss some recent results on minimax and adaptive estimation of covariance matrices in the high-dimensional setting.

The sample covariance matrix is the most commonly used estimator of the population covariance matrix in the classical fixed p, large n setting and enjoys certain optimality. When the dimension p is large, it is known that the sample covariance matrix often performs poorly. A number of regularization methods have been introduced recently for estimating large covariance matrices. Asymptotic properties and even explicit rates of convergence have been given. However, it is not clear whether any of these rates of convergence are optimal.

In this talk we begin with minimax estimation of large bandable covariance matrices. Bickel and Levina (2008a) introduced a banding estimator for estimating this class of covariance matrices and derived a rate of convergence for the banding estimator. Cai, Zhang and Zhou (2010) established the optimal rates of convergence for estimating the covariance matrix under both the operator norm and Frobenius norm. It is shown that optimal procedures under the two norms are different and consequently matrix estimation under the operator norm is fundamentally different from vector estimation. The minimax upper bound is obtained by constructing a special class of tapering estimators and by studying their risk properties. A key step in obtaining the optimal rate of convergence is the derivation of the minimax lower bound. The lower bound is established by using a testing argument, where at the core is a novel construction of a collection of least favorable multivariate normal distributions and the application of Assouad's lemma and Le Cam's method.

The rate optimal tapering estimator given in Cai, Zhang and Zhou (2010) critically depends on the parameter α which is the rate of decay of the off-diagonal entries and is thus not practical in applications. A natural goal is then to construct a single procedure which is minimax rate optimal simultaneously over each parameter space in a large collection. Cai and Yuan (2011) considered adaptive estimation of bandable matrices and proposed a block thresholding procedure. The estimator is constructed by carefully dividing the sample covariance matrix into blocks and then simultaneously estimating the entries in a block by thresholding. The estimator is shown to be optimally rate adaptive over a wide range of bandable covariance matrices.

Besides bandable covariance matrices, sparse covariance matrices also arise naturally in many applications. In this talk we also discuss minimax and adaptive estimation of sparse covariance matrices considered in Cai and Zhou (2010 and Cai and Liu (2011). In particular, Cai and Liu (2011) introduced a thresholding procedure that is adaptive to the variability of individual entries. The estimators are fully data-driven and demonstrate excellent performance both theoretically and numerically. It is shown that the estimators adaptively achieve the optimal rate of convergence over a large class of sparse covariance matrices under the spectral norm. In contrast, the commonly used universal thresholding estimators are shown to be suboptimal over the same parameter spaces.

The results and technical analysis of these high-dimensional matrix estimation problems reveal some new features that are quite different from the conventional low-dimensional or sequence estimation problems.

References

[1] Bickel, P. and Levina, E. (2008a), *Regularized estimation of large covariance matrices*, The Annals of Statistics, **36**, 199-227.

[2] Bickel, P. and Levina, E. (2008b), *Covariance regularization by thresholding*, The Annals of Statistics, **36**, 2577-2604.

[3] Cai, T.T. and Liu, W.(2011), *Adaptive thresholding for sparse covariance matrix estimation*, to appear in Journal of the American Statistical Association **106**, 672-684.

[4] Cai, T.T. and Yuan, M. (2011). *Adaptive covariance matrix estimation through block thresholding.* Technical Report.

[5] Cai, T.T., Zhang, C.-H. and Zhou, H. (2010), *Optimal rates of convergence for covariance matrix estimation*, The Annals of Statistics, **38**, 2118-2144.

[6] Cai, T. T. and Zhou, H. H. (2010). *Optimal rates of convergence for sparse covariance matrix estimation.* Technical Report.

How to analyze many contingency tables simultaneously?

Thorsten Dickhaus

Motivated by applications in the field of genetic association studies, we study exact tests for (2×2) and (2×3) contingency tables, in particular exact chi-squared tests and exact tests of Fisher-type. In practice, these tests are typically carried out without randomization, leading to reproducible results but not exhausting the significance level. We discuss that this can lead to methodological and practical issues in a multiple testing framework when many tables are simultaneously under consideration.

Realized randomized p-values are proposed as a solution which is especially suitable for usage in data-adaptive (plug-in) procedures. Although they were originally derived in terms of randomized tests, we define them in a more general way as follows.

Definition 1 *Let a statistical model $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be given. Consider the two-sided test problem $H : \{\vartheta = \vartheta_0\}$ versus $K : \{\vartheta \neq \vartheta_0\}$ and assume the decision is based on the realization $\mathbf{x}$ of a discrete random variate $\mathbf{X} \sim \mathbb{P}_\vartheta$ with values in Ω. Moreover, let U denote a uniformly distributed random variable on $[0, 1]$, stochastically independent of $\mathbf{X}$. A realized randomized p-value for testing H versus*

K is a measurable mapping $p^r : \Omega \times [0,1] \to [0,1]$ fulfilling that

$$\mathbb{P}_{\vartheta_0}(p^r(\mathbf{X}, U) \leq t) = t \text{ for all } t \in [0,1].$$

We derive convenient formulas for computing realized randomized p-values based on the chi-squared and Fisher testing strategies mentioned above. Moreover, we address the problem of positively correlated p-values for association by considering techniques to reduce multiplicity by estimating the "effective number of tests" from the correlation structure. In particular, we present and discuss the methods of Cheverud and Nyholt (cf. [1] and [5]), X. Gao et al. (see [3]) and Moskvina and Schmidt (see [4]).

An algorithm taken from [2] bundles the three aspects (i) utilization of realized randomized p-values, (ii) estimation of the proportion of true null hypotheses, (iii) estimating the effective number of tests, and we exemplify it with real data.

References

[1] J. M. Cheverud, *A simple correction for multiple comparisons in interval mapping genome scans*, Heredity **87** (2001), 52–58.

[2] T. Dickhaus, K. Straßburger, D. Schunk, C. Morcillo-Suarez, T. Illig and A. Navarro, *Refined statistical inference methods for contingency table analyses in genetic association studies*, under review.

[3] X. Gao, J. Starmer and E. R. Martin, *A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms*, Genetic Epidemiology **32** (2008), 361–369.

[4] V. Moskvina and K. M. Schmidt, *On Multiple-Testing Correction in Genome-Wide Association Studies*, Genetic Epidemiology **32** (2008), 567–573.

[5] D. R. Nyholt, *A simple correction for multiple testing for SNPs in linkage disequilibrium with each other*, Am. J. Hum. Genet. **74** (2004), 765–769.

Comments on Projection Pursuit and Empirical Processes

Lutz Dümbgen
(joint work with Perla Zerial)

Let P be a probability distribution on q-dimensional space. The so-called Diaconis-Freedman effect means that for a fixed dimension $d << q$, most d-dimensional projections of P look like a scale mixture of spherically symmetric Gaussian distributions. In this talk we present necessary and sufficient conditions for this phenomenon in a suitable asymptotic framework with increasing dimension q. It turns out that the conditions formulated by Diaconis and Freedman (1984) are not only sufficient but necessary as well. To achieve this we use a variation of a nice trick introduced already by Hoeffding (1952) in a different context.

In the second part we consider the empirical distribution $\hat{P}$ of n independent random vectors with distribution P. We investigate the behavior of the empirical process $\sqrt{n}(\hat{P} - P)$ under random projections, conditional on $\hat{P}$.

References

[1] W. Hoeffding, *The large-sample power of tests based on random permutations*, Annals of Mathematical Statistics **23** (1952), 169–192.

[2] P. Diaconis and D.A. Freedman, *Asymptotics of graphical projection pursuit*, The Annals of Statistics **12** (1984), 793–815.

[3] L. Dümbgen and P. Zerial, *On low-dimensional projections of high-dimensional distributions*, Preprint (arxiv:1107.0417).

On FDR Control, Expected Number of False Rejections and Issues under Dependence

Helmut Finner

(joint work with Veronika Gontscharuk, Sandra Landwehr, Marsel Scheer, Klaus Strassburger)

We give a brief introduction on concepts of error rate control in multiple hypotheses testing and review some recent advances and test procedures with respect to familywise error rate (FWER) and false discovery rate (FDR) control. Thereby, we restrict attention to the class of step-up-down (SUD) tests based on p-values $p_1, \ldots, p_n$ for testing n hypotheses $H_1, \ldots, H_n$. Such tests can be visualized in terms of a suitable rejection curve $r : [0, b] \longrightarrow [0, 1]$ and the empirical distribution function (ecdf) $\hat{F}_n$ of all p-values. Typically, a hypothesis H_i is rejected if $p_i \leq \tau$, where τ denotes the (random) threshold for the test procedure. Thereby, the threshold τ can be taken as one of the crossing points between r and $\hat{F}_n$, that is, τ satisfies $r(\tau) = \hat{F}_n(\tau)$. For example, for a step-up test, τ is the largest crossing point between r and $\hat{F}_n$.

We briefly review recent results for linear plug-in tests based on estimates for the number of true null hypotheses. An early reference on this issue is [11]. Meanwhile, considerable progress has been made on proving the validity of such procedures, cf. e.g. [13], [9], [3], [8], [10]. At first, we discuss finite and asymptotic ($n \to \infty$) FDR control under some basic independence assumptions (BIA) for multiple tests based on the asymptotically optimal rejection curve (AORC) introduced in [2]. Then we give a brief review concerning least favorable configurations with respect to FDR and issues appearing under dependence. We present a counterexample that even weak dependence is not sufficient to guarantee asymptotic FDR control if the so-called null-problem appears where the asymptotic threshold of a test procedure tends to 0. Weak dependence applies for example if the ecdf of p-values under true null hypotheses converges stochastically to the cdf of a uniform random variable for $n \to \infty$.

Under dependence, FWER and FDR controlling procedures often lead to a highly inflated number of false rejections, cf. e.g. [5] for FWER and [1] for FDR control. In order to bound the number of false rejections, we introduce a new and more flexible criterion for error rate control based on the expected number of false rejections (ENFR). The aim is to control the ENFR at some level function g depending on the number n_1 of false hypotheses, e. g., $g(n_1) = (n_1 + \kappa)\gamma$ for suitable constants γ, κ. Hence, conceptually similar to FDR control, we allow more false rejections if the number of false hypotheses increases. This is in contrast to [12], where ENFR control with $g \equiv \gamma$ was investigated which typically leads to conservative Bonferroni tests with $\tau = \gamma/n$. In this talk we show that there is a strong link between FDR and ENFR control under (BIA) for suitable level functions. For example, for a step-up test φ based on the critical values $\alpha_{i:n} = r^{-1}(i/n)$, $i = 1, \dots, n$, induced by the adjusted AORC $r(t) = (1 + \beta_n/n)t/[(1 - \alpha)t + \alpha]$ introduced in [2], martingale and stopping time arguments yield the striking identity

$$\mathrm{FDR}_\vartheta(\varphi) = \frac{1}{n + \beta_n}\left[(1 - \alpha)\mathrm{ENFR}_\vartheta(\varphi) + \alpha n_0\right].$$

This identity leads to the statement

$$\mathrm{FDR}_\vartheta(\varphi) = \alpha' \quad \text{iff} \quad \mathrm{ENFR}_\vartheta(\varphi) = (n_1 + \beta_n)\alpha'/(1 - \alpha').$$

Under (BIA), some further new results on the ENFR behavior of various multiple test procedures are presented, cf. e.g. [4] and [6] for some earlier results.

Very recently, A. Gordon, cf. [7], presented bounds for the ENFR under general dependence which can be utilized to construct SUD procedures controlling the ENFR. However, such procedures are typically very conservative.

Finally, we illustrate dependency issues arising in gene expression data from the KORA study (a cooperation of the German Diabetes Center Düsseldorf with the Helmholtz Center Munich). Clearly, in comparing the expression levels of about 32000 gene transcripts between different groups of individuals (e.g. diabetics versus non-diabetics), p-values cannot be assumed to be independent. Among others, we illustrate the null-problem and investigate whether weak dependence may apply by studying the correlation structure in the data.

References

[1] H. Finner, T. Dickhaus, M. Roters, *Dependency and false discovery rate: Asymptotics*, Ann. Statist. **35** (2007), 1432–1455.

[2] H. Finner, T. Dickhaus, M. Roters, *On the false discovery rate and an asymptotically optimal rejection curve*, Ann. Statist. **37** (2009), 596–618.

[3] H. Finner, V. Gontscharuk, *Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses*, J. Roy. Statist. Soc. **B 71** (2009), 1031–1048.

[4] H. Finner, M. Roters, *On the false discovery rate and expected type I errors*, Biom. J. **43** (2001), 985–1005.

[5] H. Finner, M. Roters, *Asymptotic sharpness of product-type inequalities for maxima of random variables with applications in multiple comparisons*, J. Statist. Plann. Inf. **98** (2001), 35–56.

[6] H. Finner, M. Roters, *Multiple hypotheses testing and expected number of type I errors*, Ann. Statist. **30** (2002), 220–238.

[7] A. Y. Gordon, *A sharp upper bound for the expected number of false rejections*, *Talk at "The 7th International Conference on Multiple Comparison Procedures (MCP 2011), Washington D. C.* (2011), `http://www.mcp-conference.org/2011/files/final_program.pdf`.

[8] W. Guo, *A note on adaptive Bonferroni and Holm procedures under dependence*, Biometrika **96** (2009), 1012–1018.

[9] S. K. Sarkar, *On methods controlling the false discovery rate (with discussions)*, Sankhya **70** (2008), 135–168.

[10] S. K. Sarkar, W. Guo, H. Finner, *On adaptive procedures controlling the familywise error rate*, J. Statist. Plann. Inf. **142** (2012), 65–78.

[11] T. Schweder, E. Spjøtvoll, *Plots of p-values to evaluate many tests simultaneously*, Biometrika **69** (1982), 493–502.

[12] E. Spjøtvoll, *On the optimality of some multiple comparison procedures*, Ann. Statist. **43** (1972), 398–411.

[13] J. D. Storey, J. E. Taylor, D. Siegmund, *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach.* J. R. Stat. Soc. **B** 66 (2004), 187–205.

Fast Bayesian model assessment for nonparametric additive regression with many predictors

Subhashis Ghoshal

(joint work with S. Mckay Curtis, Sayantan Banerjee)

We consider a Bayesian approach for nonparametric additive regression model with many predictors. We expand the functions in the additive model in a B-spline basis and put a mixture of point mass at zero vector and a multivariate Laplace prior on coefficients of each basis function. The indicators of non-null coefficients are given a suitable prior based on the dependence structure of the predictor variables. It is observed that conditionally on these indicators, the posterior mode can be viewed as a group LASSO estimator. The posterior density can be expanded around the posterior mode to yield Laplace approximations of posterior probabilities of various submodels generated by predictor selection. This leads to extremely fast assessment of posterior probabilities of various submodels of interest, and can be used for model averaging and prediction.

References

[1] Yuan, M. and Lin, Y. *Efficient empirical Bayes variable selection and estimation in linear models*, Journal of the American Statistical Association **100** (2005), 1215–1225.

[2] Yuan, M. and Lin, Y. *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B **68** (2006), 48–57.

Penalized estimation of high dimensional models under a generalized sparsity condition

Joel Horowitz

This talk is about estimation of a linear or nonparametric additive model in which the number of regression coefficients or additive components may exceed the sample size. Motivated by applications in economics, we assume that a few coefficients or additive components are large and objects of interest, whereas many others are small but not necessarily zero. The large coefficients or additive components can be estimated more accurately if the small ones can be identified and the covariates associated with them dropped from the model. We show that this can be done with a two-step procedure in which the first step is the LASSO or group LASSO and the second step is a form of penalized least squares. Monte Carlo experiments and an empirical application illustrate the usefulness of the procedure.

UPS delivers optimal phase diagram in high dimensional variable selection

Jiashun Jin
(joint work with Pengsheng Ji)

Consider a linear regression model

$$Y = X\beta + z, \qquad z \sim N(0, I_n), \qquad X = X_{n,p},$$

where both p and n are large but $p > n$. The vector β is unknown but is sparse in the sense that only a small proportion of its coordinates is nonzero, and we are interested in identifying these nonzero ones. We model the coordinates of β as samples from a two-component mixture $(1-\epsilon)\nu_0 + \epsilon\pi$, and the rows of X as samples from $N(0, \frac{1}{n}\Omega)$, where ν_0 is the point mass at 0, π is a distribution, and Ω is a p by p correlation matrix which is unknown but is presumably sparse.

We propose a two-stage variable selection procedure which we call the *UPS*. This is a Screen and Clean procedure, in which we screen with the Univariate thresholding, and clean with the Penalized MLE. In many situations, the UPS possesses two important properties: Sure Screening and Separable After Screening (SAS). These properties enable us to reduce the original regression problem to many small-size regression problems that can be fitted separately. As a result, the UPS is effective both in theory and in computation.

We measure the performance of variable selection procedure by the Hamming distance, and use an asymptotic framework where $p \to \infty$ and $(\epsilon, \pi, n, \Omega)$ depend on p. We find that in many situations, the UPS achieves the optimal rate of convergence. We also find that in the (ϵ_p, π_p) space, there is a three-phase diagram shared by many choices of Ω. In the first phase, it is possible to recover all signals. In the second phase, exact recovery is impossible, but it is possible to recover most of the signals. In the third phase, successful variable selection is impossible. The UPS partitions the phase space in the same way that the optimal procedures do, and recovers most of the signals as long as successful variable selection is possible.

The lasso and the subset selection (also known as the L^1- and L^0-penalization methods, respectively) are well-known approaches to variable selection. However, somewhat surprisingly, there are regions in the phase space where neither the lasso nor the subset selection is rate optimal, even for very simple Ω. The lasso is non-optimal because it is too loose in filtering out fake signals (i.e. noise that is highly correlated with a signal), and the subset selection is non-optimal because it tends to kill one or more signals in correlated pairs, triplets, etc..

Testing monotonicity of a hazard rate

Geurt Jongbloed
(joint work with Piet Groeneboom)

In reliability theory and survival analysis, the hazard rate h is a natural function to characterize the distribution of a nonnegative random variable X with probability density function f. It is defined by

$$h(x) = \frac{f(x)}{1 - F(x)},$$

where F is the distribution function of X. If X describes the time a typical electrical component from a batch can be effectively used, monotonicity properties of h indicate whether the components in the batch become more / less reliable during a time period in which these are used (decreasing / increasing h on time interval).

Let $X_1, X_2, \ldots$ be i.i.d. random variables with hazard rate h. In this presentation, the problem of testing the null hypothesis that h is increasing on a certain time interval will be considered, based on $X_1, \ldots, X_n$. An L_1-type test statistic based on an L_2-projection estimator will be introduced. The asymptotic distribution of this statistic will be given, as well as an outline of its derivation. Moreover, a practical bootstrap-based procedure will be described and simulation results will be presented, comparing the procedure with natural competitors introduced in the literature ([1],[4]). The talk is based on [2] and [3].

References

[1] C. Durot, *Testing convexity or concavity of a cumulated hazard rate*, IEEE Transactions on Reliability **57** (2008), 465 - 473.

[2] P. Groeneboom and G. Jongbloed, *Testing monotonicity of a hazard: asymptotic distribution theory*, Submitted

[3] P. Groeneboom and G. Jongbloed, *Isotonic L_2-projection test for local monotonicity of a hazard*, Submitted

[4] P. Hall and I. van Keilegom, *Testing for monotone increasing hazard rate.* The Annals of Statistics **33** (2005), 1109-1137.

Occupancy problems in high-dimensional space

Estate Khmaladze

A. Consider a block of unit mass which we split into k smaller blocks of sizes, following some distribution F; each of these smaller blocks are again split in the same manner, each into k smaller blocks; etc..

After d steps we will have $N = k^d$ small pieces of sizes $\xi_{1d}, \ldots, \xi_{Nd}$. Consider empirical distribution function of "magnified" sizes

$$H_N(z) = \frac{1}{N} \sum_{i=1}^{N} 1_{\{N\xi_{id} > z\}}.$$

For a very wide class of splitting distributions F, we show that $H_N(z) \to 0$ for every $z > 0$, and moreover, as $d \to \infty$,

$$H_N(z) \sim C_n \frac{1}{z^u}, \quad 0 < u < 1, \quad C_n \to 0. \tag{1}$$

B. Suppose each particle emitts (or is occupied by) a Poisson process with intensity $N\xi_{id}$. For a fixed time t, let $\mu_t(k)$ be the number of these Poisson processes, which are equal k and μ_t be a number of Poisson processes which are positive.

Then we show that, as $d \to \infty$,

$$\frac{E\mu_t(k)}{E\mu_t} \longrightarrow \frac{u}{\Gamma(1-u)} \frac{\Gamma(k-u)}{\Gamma(k+1)} \tag{2}$$

and the limit does not depend on t. The convergence in (2) is a corollary of the convergence in (1).

Aspects of the Bernstein-von Mises theorem

Bas Kleijn

We consider a LAN model $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta \subset \mathbb{R}^k, \eta \in H\}$ where H is ∞-dimensional with a prior Π and ask whether the θ-posterior converges to a normal limit distribution $N(\hat{\theta}_n, n^{-1}\tilde{I}_0^{-1})$ where $\hat{\theta}_n$ is the MLE and $\tilde{I}_0^{-1}$ is the efficient Fisher information. If the prior $\Pi = \Pi_\Theta \times \Pi_H$ satisfies Schwartz's consistency conditions (Schwartz (1965)) and the marginal posterior for θ converges at $n^{-1/2}$-rate, the Bernstein-von Mises limit holds. The latter condition can only be satisfied if the bias is controlled. We propose to "regularize" the prior by introduction of a density of the form e^{I_n}, where $I_n : H \to R$ is the penalty one would add to the log-likelihood in penalized MLE procedures known from point-estimation, as in van de Geer (2000).

Modelling extremes observed in space and time

Claudia Klüppelberg

(joint work with Richard A. Davis and Christina Steinkohl)

Often, in modelling complex systems such as wind fields, statistical methodology can be applied to reconcile the physical models with the data. For an adequate risk analysis, the statistical modelling of extreme events, such as severe wind gusts or storms is essential; cf. [4].

Historically, Gaussian random fields play a central role in modelling spatio-temporal data. When it comes to extremes and extremal dependence, Gaussian processes are not appropriate, since observations at two different locations and time points are in Gaussian models independent at high levels. A natural extension from uni- and multivariate extreme value theory is formed by so-called max-stable random fields. We suggest new statistical models for extreme data measured in space and time. We present the basic aspects and challenges of simulation and estimation of max-stable spatio-temporal processes.

Our simulation method is based on limit results for Gaussian models derived originally by [3] for bivariate models. It requires a certain limiting property of the correlation structure, which extends to the much more general space-time setting; see also [7]. Our construction extends also [8], who works in a pure spatial setting.

We calculate the bivariate distribution functions of the max-stable spatio- temporal process, which involves the underlying Gaussian dependence structure. Estimation can now be based on the pairwise likelihood; cf. also [5]. Finally, we prove strong consistency of the pairwise maximum likelihood estimators.

We test our procedure by first simulating a specific max-stable spatio-temporal process, based on an underlying Gaussian field with a correlation function of the Gneiting class [2]. We use our pairwise likelihood estimation, which turns out to work very well in practice.

References

[1] R.A. Davis, C. Klüppelberg and C. Steinkohl (2011) *Max-stable processes for modelling extremes observed in space and time.* Submitted.

[2] Gneiting, T. (2002) *Nonseparable, stationary covariance functions for space-time data.* JASA **95**, 590-600.

[3] J. Hüsler and R.-D. Reiss (1989) *Maxima of normal random vectors: between independence and complete dependence.* Statistics and Probability Letters **7**, 283-286.

[4] M. Nielsen, G.C. Larsen, J. Mann, S. Ott, K.S. Hansen and B.J. Pedersen (2004) *Wind simulation for extreme and fatigue loads.* Risø National Laboratory, Denmark.

[5] S.A. Padoan and M. Ribatet and S.A. Sisson (2009) *Likelihood-based inference for max-stable processes.* JASA **105**, 263-277.

[6] M. Schlather (2002) *Models for stationary max-stable random fields.* Extremes **5**(1), 33-44.

[7] Z. Kabluchko, M. Schlather and L. de Haan (2009) *Stationary max-stable fields associated to negative definite functions.* Ann. Probab. **37**(5), 2042 - 2065.

[8] R. L. Smith (1990) *Max-stable processes and spatial extremes.* Unpublished manuscript

A Framework For Estimating Convex Functions

Mark Low

The problem of estimating a function assumed to be convex is an important special case of a large collection of problems focused on estimation under order constraints. We consider a white noise with drift model where the drift is assumed to be convex and focus on estimating the drift function at a given point under squared error loss. The main goal is to develop a general framework for the evaluation of specific procedures. This framework develops a benchmark tied to each convex function by considering the hardest alternative for this function. A local modulus of continuity is introduced to express this benchmark. The lower bound is also expressed in an easily computable way through the introduction of a new function which we term the k function. A procedure is constructed which is almost optimal for each function.

Integrative analysis of cancer genomic studies

Shuangge Ma

(joint work with Jian Huang)

In high-throughput cancer prognosis studies, markers identified from the analysis of single datasets often suffer a lack of reproducibility because of the small sample sizes. A cost-effective remedy is to pool data from multiple existing studies and conduct integrative analysis. In this study, we describe cancer survival using AFT (accelerated failure time) models. A weighted least squared criterion is proposed for estimation. We propose a group MCP penalization approach for integrative analysis of multiple heterogeneous prognosis studies and marker selection. We establish the asymptotic selection consistency properties under the condition $p = \exp(o(n))$, where p is the number of covariates and n is the combined sample size. Simulations and analysis of breast cancer studies show that the proposed approach outperforms individual-dataset analysis and meta-analysis.

Asymptotic equivalence of functional linear regression and a white noise inverse problem

Alexander Meister

We consider the statistical experiment of functional linear regression (FLR) under normal regression errors. Furthermore, we introduce a white noise model where one observes an Ito process, which contains the covariance operator of the corresponding FLR model in its construction. We prove asymptotic equivalence of FLR and this white noise model in LeCam's sense under known design distribution. Moreover, we show equivalence of FLR and an empirical version of the white noise model for finite sample sizes. As an application, we derive sharp minimax constants in the FLR model which are still valid in the case of unknown design distribution. The talk is based on the paper [1].

References

[1] A. Meister, *Asymptotic equivalence of functional linear regression and a white noise inverse problem*, The Annals of Statistics **39** (2011), 1471–1495.

Statistical Multiscale Methods and Biophotonic Imaging

Axel Munk

(joint work with K. Frick, Z. Kabluchko, P. Marnitz, H. Sieling, A. Egner, S. Hell, A. Schönle)

A central goal in statistical signal detection and imaging is to recover an unknown (gray-valued) signal/image u defined on some domain $\Omega \subset \mathbb{R}^d$ from data Y. We shall denote the collection of all images on Ω by U and assume that U is a linear space. A common model which serves as a proxy for many practical situations is that the measurement takes place on some finite grid $X = \{1, \dots, m\}^d$ of size n and that for each multi-index $\nu \in X$

$$Y_\nu = (Ku)_\nu + \varepsilon_\nu, \tag{1}$$

where ε_ν are i.i.d. centered Gaussian random variables with presumably known variance $\sigma^2 > 0$ and $K : U \to \mathbb{R}^n$ is a linear operator that encodes the functional relation between the quantities that are accessible by experiment and the underlying signal/image. A typical example for K is convolution, leading to a blurred version Ku of u.

In practical situations most of the images contain features of different scales and modality, i.e. constant and smooth portions as well as oscillating patterns both of different sizes. Thus, a minimum claim for modern reconstruction methods is to allow for such spatially varying characteristics. In the recent literature on such methods one can roughly distinguish between two approaches

- **Sparse Dictionary based Multiscale Methods** aim for representing an estimator $\hat{u} \in U$ w.r.t. some given dictionary of "localizing" functions and determine the corresponding coefficients according to the extreme value behaviour of the residual values $Y - K\hat{u}$ (thresholding methods) or refined risk calculations [1]. Standard examples are wavelet- or curvelet dictionaries (see [2, 7] among many others).
- **Variational Methods** typically compute estimators as minimizers of penalized likelihood functions, where the penalty encodes a priori knowledge on the regularity (smoothness, sparsity, ...) of the unknown function u. Prominent examples are total-variation based methods ([11, 5, 4])

  $$\hat{u} = \operatorname*{argmin}_{u \in U} \sum_{\nu \in X} (Ku - Y)_\nu^2 + \lambda \int_\Omega |\nabla u|.$$

  Here the parameter $\lambda > 0$ governs the trade-off between data-fit and smoothness and it is claimed that locality is expressed through $\int_\Omega |\nabla u|$.

In the following we aim for unifying these to seemingly different approaches. To this end we introduce the following class of estimators. Let $\mathcal{S}$ be some index set and $\mathcal{W} = \{\omega^S,\ S \in \mathcal{S}\}$ be a set of given weight-functions on the grid X. A *statistical multiresolution estimator (SMRE)*, is defined as a solution of the constrained optimization problem

$$J(u) \to \inf! \quad \text{s.t.} \quad \max_{S \in \mathcal{S}} \left| \sum_{\nu \in X} \omega_\nu^S \left(\Lambda(Y - Ku)\right)_\nu \right| \leq q. \tag{2}$$

Here, $J : U \to \mathbb{R}$ denotes a regularization functional that incorporates a priori knowledge on the unknown true signal u^0 (such as smoothness, sparsity or texture information) and $\Lambda : \mathbb{R}^n \to \mathbb{R}^n$ a possibly non-linear transformation. The constant q can be considered as a *regularization parameter* that governs the trade-off between regularity and data-fit of the reconstruction. In most practical situations q is chosen to be the (asymptotic) α-quantile q_α of the *multiresolution (MR) statistic*

$T(\epsilon)$, where T satisfies the inequality constraint in (2) and is defined as

$$T(v) = \max_{S \in \mathcal{S}} \left| \sum_{\nu \in X} \omega_{\nu}^{S} \left(\Lambda(v)\right)_{\nu} \right|, \quad v \in (\mathbb{R}^m)^d. \tag{3}$$

The regularization parameter q admits a universal and sound statistical interpretation: each solution $\hat{u}_\alpha$ of (2) satisfies

$$\mathbb{P}\left(J(\hat{u}_\alpha) \leq J(u^0)\right) \geq \alpha$$

i.e. the estimator $\hat{u}_\alpha$ is *more regular* (in terms of J) than u^0 with a probability of at least α.

For a given estimator $\hat{u}$ of u^0, the set $\mathcal{W}$ is assumed to be rich enough in order to catch all relevant non-random signals that are visible in the residual $Y - K\hat{u}$. Put differently, the MR-statistic $T(Y - K\hat{u})$ is bounded by q, whenever $Y - K\hat{u}$ is accepted as white noise according to the *resolution* provided by $\mathcal{W}$.

Summarizing, the optimization problem in (2) amounts to choose the most parsimonious among all estimators $\hat{u}$ for which the residual $Y - K\hat{u}$ resembles white noise according to the statistic T. If $Y - K\hat{u}$ contains some non random signal, $T(Y - K\hat{u})$ is likely to be larger than q and u happens to lie outside the admissible domain of (2). Thus, the multi-resolution constraint prevents too parsimonious reconstructions due to the minimization of J.

In fact, this general SMRE approach is applicable to numerous application areas. The examples we have primarily in mind are mainly from *signal detection* and *statistical imaging*, where the index set $\mathcal{S}$ is often chosen to be an overlapping (redundant) system of subsets of the grid X and ω^S is the normalized indicator function on S.

Therefore, we obtain a reconstruction method that can be combined with any variational functional J and *locally adapts the amount of regularization* according to the underlying signal or image features.

The multiscale approach in conjunction with variational regularization has been advocated for dimension $d = 1$ by several authors, see e.g. [8] for testing for local monotonicity and [4] for the case of total variation (see also [6]). In [9] this is extended to general convex functionals J, and for higher dimensions ($d \geq 2$) as well as to deconvolution problems. To this end an algorithmic framework is presented which allows to decompose the minimization problem in (2) into an unconstrained mimimization of J and a separate projection step onto the convex multiresolution constraint by an augmented Lagrange approach. This reveals these optimization problems as computationally feasible even in large scale deconvolution problems as they arise in nanoscale biophotonic imaging [10].

Finally, this also generalizes prominent regularization techniques in high dimensional statistics to multi scales, including the Danzig selector [3], which originally has been defined on a single scale.

References

[1] Cai, T., Zhou, H. (2009). *A data-driven block thresholding approach to wavelet estimation.* The Annals of Statistics 37, 569-595.

[2] Candès, E. J., Donoho. D. L. (2004). *New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities.* Comm. Pure Appl. Math. 57, 219–266.

[3] Candès, E., Tao, T. (2007). *The Dantzig selector: statistical estimation when p is much larger than n.* The Annals of Statistics 35, 2313 – 2351.

[4] Davies, P. L., Kovac, A. (2001). *Local extremes, runs, strings and multiresolution.* Ann. Statist. 29, 1 – 65. With discussion and rejoinder by the authors.

[5] Dobson, D. C., Vogel, C. R. (1997). *Convergence of an iterative method for total variation denoising.* SIAM J. Numer. Anal. 34, 1779 – 1791.

[6] Dong, Y., Hintermüller, M., Rincon-Camacho, M. (2011). *Automated regularization parameter selection in multi-scale total variation models for image restoration.* J. Math. Imaging Vision 40, 82 – 104.

[7] Donoho, D. L. (1993). *Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data.* In: Different Perspectives on Wavelets, volume 47 of Proc. Sympos. Appl. Math., pages 173 – 205, Providence, RI, 1993. Amer. Math. Soc.

[8] Dümbgen, L., Spokoiny, V. G. (2001). *Multiscale testing of qualitative hypotheses.* The Annals of Statistics 29, 124 – 152.

[9] Frick, K., Marnitz, P., Munk, A., (2011). *Statistical multiresolution estimation in imaging: Fundamental concepts and algorithmic framework.* ArXiv: 1101.4373

[10] Hell, S. W. (2007). Far-field optical nanoscopy. *Science* 316, 1153 – 1158.

[11] Rudin, L.I., Osher, S., Fatemi, E., (1992). *Nonlinear total variation based noise removal algorithms.* Phys. D 60, 259 – 268.

Finite Approximation of VC Classes

Andrew B. Nobel

(joint work with Terrence M. Adams)

Let $(\mathcal{X}, \mathcal{S}, \mu)$ be a probability space and let $\mathcal{C} \subseteq \mathcal{S}$ be a given family of measurable sets. The Vapnik-Chervonenkis dimension of $\mathcal{C}$ is a measure of its combinatorial complexity. Given a finite set $D \subseteq \mathcal{X}$, let $\{C \cap D : C \in \mathcal{C}\}$ be the collection of subsets of D selected by the members of $\mathcal{C}$. The family $\mathcal{C}$ is said to shatter D if its elements can select every subset of D, or equivalently, if $|\{C \cap D : C \in \mathcal{C}\}| = 2^{|D|}$. The Vapnik-Chervonenkis (VC) dimension of $\mathcal{C}$, denoted $\dim(\mathcal{C})$, is the largest integer k such that $\mathcal{C}$ is able to shatter *some* set of cardinality k. If $\mathcal{C}$ can shatter arbitrarily large finite sets, then the dimension of $\mathcal{C}$ is infinite.

Let π be a finite, measurable partition of $\mathcal{X}$. For every set $C \in \mathcal{C}$, the π-boundary of C, denoted $\partial(C : \pi)$, is the union of all the cells in π that intersect both C and its complement with positive probability. Formally,

$$\partial(C : \pi) \;=\; \cup\{A \in \pi : \mu(A \cap C) \cdot \mu(A \cap C^c) > 0\}.$$

Note that $\partial(C : \pi)$ depends on μ, though this dependence is suppressed in our notation. We will call a family $\mathcal{C}$ *finitely approximable* if for every $\epsilon > 0$ there exists a finite, measurable partition π of $\mathcal{X}$ such that $\mu(\partial(C : \pi)) \leq \epsilon$ for every $C \in \mathcal{C}$. The principal subject of the talk, established in [1], is the following.

Theorem: If $\mathcal{C}$ has finite VC dimension, then $\mathcal{C}$ is finitely approximable for every probability measure μ on $(\mathcal{X}, \mathcal{S})$.

The theorem does not impose any cardinality or regularity constraints on the family of sets $\mathcal{C}$ beyond the purely combinatorial requirement that $\mathcal{C}$ have finite VC dimension. Immediate corollaries of the theorem include the fact that pointwise separable classes with finite VC dimension have finite bracketing numbers, and satisfy uniform laws of large numbers for every ergodic process. Details, and extensions to families of real-valued processes, can be found in [1].

References

[1] Adams, T.M. and Nobel, A.B. *Uniform approximation and bracketing properties of VC classes*, to appear in Bernoulli (2011).

Estimation of the Lévy measure: statistical inverse problem and a Donsker theorem

Markus Reiss
(joint work with Richard Nickl)

Given n equidistant observations of a Lévy process $(L_t,\ t \geq 0)$ with Lévy measure ν we construct estimators $\hat{\nu}_n$ of ν and assess their performance by looking at integrals $\int f d\nu$ for integrands f of smoothness s. The nonlinear estimator attains rates as for estimating regular functionals in statistical inverse problems. The ill-posedness is – like in deconvolution – prescribed by the decay of the characteristic function ϕ (joint work with Michael Neumann, Jena [1]). In a second step we construct a natural estimator $\hat{N}_n$ of the generalised distribution function $N(t) = \nu((-\infty, t])$ for $t < 0$ such that $\sqrt{n}(\hat{N}_n - N)$ satisfies a functional central limit theorem in the space of bounded functions. The limit distribution is a generalised Brownian bridge with a covariance structure that is minimal in the Cramér-Rao sense. The class of Lévy processes covered includes several concrete examples, such as compound Poisson, Gamma and self-decomposable processes whose characteristic functions obey a natural decay restriction. Mathematical tools include pseudo-differential operator calculus and smoothed empirical processes.

References

[1] M.H. Neumann and M.Reiss *Nonparametric estimation for Levy processes from low-frequency observations* Bernoulli **15** (1), (2009), 223–248.

Bayesian non-parametric estimators in high dimension

Ya'acov Ritov
(joint work with Peter J. Bickel, Anthony Gamst)

In this talk we considered the possibility of having a honestly non-parametric Bayesian procedure in the context of a very high model. We considered this problem in the context of the white noise model with heavy tails, $Y_i = \beta_i + \epsilon_i$, $\sum i^{2\alpha}\beta_i^2 < \infty$, $\alpha \in (1/4, 1/2)$. We show that any Bayesian procedure under quadratic loss function, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots)$, is going to fail to achieve the efficient convergence rate either as a non-parametric estimator of $\beta = (\beta_1, \beta_2, \dots)$, or as a parametric estimator of some functional $h(\beta) = \sum h_i\beta_i$, with $\sum h_i^2 = 1$. We argue that in some sense, this would happen for most relevant functionals h, and this is follow that any non-parametric Bayes estimator of β is necessarily biased, and hence this bias would make the estimator of $h(\beta)$ biased as well. We argue that in many situations, a Bayes estimator can be efficient both as a non-parametric estimator and as a parametric estimator of *specific* functionals, but this can be achieved only if the prior is tuned to these functionals (and hence is not a subjective prior).

Minimax Inference Using Higher Order Influence Functions

James Robins
(joint work with Lingling Li, Eric Tchetgen, Aad van der Vaart)

Perhaps the most common model used in analyzing observational studies of the causal effect of a binary treatment A on a continuous response Y, in the presence of a vector X of continuous pretreatment confounding variables is the semiparametric regression model

$$E\left[Y|A,X\right] = \beta A + \nu(X), \tag{1}$$

where β is an unknown parameter and $\nu(\cdot)$ is an unknown function of X. This model arises whenever we assume (i) no unmeasured confounders (i.e., ignorability of treatment A within levels of X) and (ii) a constant additive effect of treatment A on the mean of Y. The parameter β in model (1) equals the ratio

$$\tau \equiv E\left[cov(Y,A|X)\right]/E\left[var(A|X)\right]$$

of the expected conditional covariance of Y and A to the expected conditional variance of A. Model (1) assumes that the treatment effect function

$$\gamma(x) = E(Y|A=1,X=x) - E(Y|A=0,X=x)$$

does not depend on x. When model (1) is misspecified and thus β is undefined, many semiparametric estimators of β continue to converge in probability to τ.

Therefore, we will study nonparametric point and interval estimation of the functional τ, both with and without imposing the often unrealistic assumption that model (1) holds. When A is binary, τ is also the variance-weighted average treatment effect functional $E\left[var(A|X)\gamma\left(X\right)\right]/E\left[var(A|X)\right]$.

For any $\tau^{*}\in R$, it is useful to define $Y\left(\tau^{*}\right)=Y-\tau^{*}A$ and the corresponding functional

$$\psi\left(\tau^{*}\right)=E\left[\left\{Y\left(\tau^{*}\right)-E\left[Y\left(\tau^{*}\right)|X\right]\right\}\left\{A-E\left(A|X\right)\right\}\right].$$

$\psi\left(\tau^{*}\right)$ is of interest because $\tau=E\left[cov(Y,A|X)\right]/E\left[var(A|X)\right]$ is the unique solution to the equation $\psi\left(\tau^{*}\right)=0$. Thus inference on τ is easily obtained from inference on $\psi\left(\tau^{*}\right)$. In particular a $(1-\alpha)$ confidence set for τ is the set of values taken by τ^{*} such that a $(1-\alpha)$ confidence interval for $\psi\left(\tau^{*}\right)$ contains zero. Optimality results we obtain for our proposed estimators $\widehat{\psi}\left(\tau^{*}\right)$ of $\psi\left(\tau^{*}\right)$ extend to the estimators $\widehat{\tau}$ of τ satisfying $\widehat{\psi}\left(\widehat{\tau}\right)=0$.

Thus our goal is to construct point and interval estimators of the functional $\psi\left(\tau^{*}\right)$ for a fixed τ^{*}. Moreover, without loss of generality, we can take $\tau^{*}=0$, so our statistical problem becomes point and interval estimation of the parameter of interest, the expected conditional covariance $\psi\left(0\right)\equiv\psi\equiv E\left[cov\left\{Y,A|X\right\}\right]$. Note that ψ is the expected conditional variance of Y in the special case that Y and A are equal with probability one.

A number of authors have considered nonparametric estimation of expected conditional covariances and/or variances. The paper by Cai et al. (2009) is most relevant.

We assume we observe N iid copies of $O=(Y,A,X)$ sampled from a probability measure F_{O} with X a $d-$dimensional vector with compact support in R^{d} and marginal density $g\left(x\right)$ that is absolutely continuous on its support wrt to Lebesgue measure. Without loss of generality, we take the support of X to be the unit hypercube $\left[0,1\right]^{d}$. We assume F_{O} is contained in a non or semi-parametric model $M\left(\Theta\right)=\left\{F(\cdot;\theta);\theta\in\Theta\right\}$, indexed by the parameter $\theta\in\Theta$. In the following, $b:x\mapsto b\left(x\right)=E\left[Y|X=x\right]$, $p:x\mapsto p\left(x\right)=E\left[A|X=x\right]$, and $g:x\mapsto g\left(x\right)$ denote the components of θ corresponding to the conditional expectations of Y and A given $X=x$ and the density of the marginal distribution F_{X} of X.

Denote the expected conditional covariance functional by

$$\psi\left(\theta\right)=E_{\theta}\left[cov_{\theta}(Y,A|X)\right].$$

Assume:

Each $h\in\{b,p,g\}$ belongs to a Hölder class of smooth functions $H(\beta_{h},C_{h})$ with known Hölder exponent β_{h} and radius C_{h}.

We have proposed a novel class of estimators for $\psi\left(\theta\right)$. Our novel estimators are higher order U-statistics. They are based on the theory of higher order influence functions for smooth nonlinear functionals in high dimensional semi/nonparametric models introduced in Robins et al. (2008).

In our case the semiparametric model is the model that assumes $cov_{\theta}(Y,A|X)$ does not depend on X.

Define

$$\beta=\left(\beta_{p}+\beta_{b}\right)/2$$

Results for the "regular" case are:

Theorem

(i) If $\beta/d \geq 1/4$ then minimax rate is $n^{1/2}$ for estimating $\psi(\theta)$ in both semi and nonparametric models even if no smoothness on g is assumed.

(ii) If $\beta/d > 1/4$, efficient (regular, AL, attains SVB) adaptive estimator available in either non or semiparametric model regardless of smoothness of g.

We are more interested in the irregular case $\beta/d < 1/4$. For simplicity assume $\beta_p = \beta_b = \beta$. Note

$$\begin{aligned} n^{-1/2} &< n^{-\frac{4\beta/d}{1+4\beta/d}} < n^{-2\beta/d} \textit{ when } \beta/d < 1/4 \\ n^{-1/2} &> n^{-\frac{4\beta/d}{1+4\beta/d}} \text{ when } \beta/d > 1/4, \end{aligned}$$

with equality at $\beta/d = 1/4$. The following table summarizes our results and conjectures concerning optimal rates of convergence for either of the functionals $E_\theta\left[Cov_\theta\left\{Y, A|X\right\}\right]$ or $E_\theta\left[var_\theta\left\{Y|X\right\}\right]$

	NP	SP
X Non-Random, ES	$n^{-2\beta/d}(U+L)$	$n^{-2\beta/d}(U+L)$
X Random, g=AC and known_	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+L)$	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+L^*)$
X Random, $\frac{\beta_g}{d} > \frac{\{\beta/d\}(1-4\beta/d)}{1+2\beta/d\ +8(\beta/d)^2}$	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+L)$	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+L^*)$
X Random, g=AC, $\beta \leq 1$	$n^{-2\beta/d}(U+L^*)$	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+\mathrm{L}^*)$
X Random, g some rate, $\beta > 1$	$n^{-2\beta/d}(U+L^*)$	$n^{-\frac{4\beta/d}{1+4\beta/d}}(U+\mathrm{L}^*)$

where:
g=density of X
AC: absolutely continuous wrt uniform measure on the d-dimensional unit cube
ES: equally spaced design.
SP model assumes $var_\theta\left\{Y|X\right\}$ or $var_\theta\left\{Y|X\right\}$ do not depend on X
U an upper bound on the rate obtained by Cai et al (2009) in the nonrandom case and with our higher order U-statistic estimators in the random case.
L a lower bound proved by Cai et al. in the nonrandom case and Robins et al in the random case.
L^* means a lower bound that is conjectured but not proved.

Alternation and semiparametric efficiency

Vladimir Spokoiny

The talk discusses the problem of efficiency for sequential procedures in semiparametric estimation. The use of the 'modern' version of the Le Cam's theory of statistical experiments allows to reduce this problem to be a linear one. We show that under weak identificability conditions, linear alternation leads to the efficient procedure. This allows to state similar results in general situations under some regularity conditions.

Matrix Uncertainty Selector under Random Noise in the Matrix

Alexandre B. Tsybakov
(joint work with Mathieu Rosenbaum)

We consider the regression model with observation error in the design:

$$\begin{aligned} y &= X\theta^* + \xi, \\ Z &= X + \Xi. \end{aligned}$$

Here the random vector $y \in \mathbb{R}^n$ and the random $n \times p$ matrix Z are observed, the $n \times p$ matrix X is unknown, Ξ is an $n \times p$ random noise matrix, $\xi \in \mathbb{R}^n$ is a random noise vector, and θ^* is a vector of unknown parameters to be estimated. We consider the setting where the dimension p can be much larger than the sample size n and θ^* is sparse. For example, the case where the entries of the matrix X are missing at random can be boiled down to this model.

It has been shown in Rosenbaum and Tsybakov (2010) that the presence of observation noise has severe consequences on the usual estimation procedures in the high-dimensional setting. In particular, the Lasso and Dantzig selector turn out to be inaccurate and fail to identify the sparsity pattern of the vector θ^*.

In the same paper, the authors provide an alternative procedure, called Matrix Uncertainty selector (MU selector for short), which is robust to the presence of noise. The MU selector $\hat{\theta}^{MU}$ is defined as a solution of the minimization problem

$$\min\{|\theta|_1 : \ \theta \in \Theta, \ \Big|\frac{1}{n} Z^T(y - Z\theta)\Big|_\infty \le \mu|\theta|_1 + \tau\},$$

where $|\cdot|_p$ denotes the ℓ_p-norm, $1 \le p \le \infty$, Θ is a given subset of $\mathbb{R}^p$ characterizing the prior knowledge about θ^*, and the constants μ and τ depend on the level of the noises Ξ and ξ respectively.

Here we suggest a modification of the MU selector in the case where Ξ is a random matrix with independent and zero mean entries Ξ_{ij} such that the sums of expectations

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\Xi_{ij})$$

are finite and admit data-driven estimators. This is for example the case in the model with missing data:

$$\tilde{Z}_{ij} = X_{ij}\eta_{ij}, \quad i = 1, \dots, n, \ j = 1, \dots, p,$$

where for each fixed $j = 1, \dots, p$, the factors $\eta_{ij}, i = 1, \dots, n$, are i.i.d. Bernoulli random variables taking value 1 with probability $1 - \pi_j$ and 0 with probability π_j, $0 < \pi_j < 1$. This model can indeed be rewritten under the form

$$Z_{ij} = X_{ij} + \Xi_{ij},$$

where $Z_{ij} = \tilde{Z}_{ij}/(1-\pi_j)$ and $\Xi_{ij} = X_{ij}(\eta_{ij} - (1-\pi_j))/(1-\pi_j)$. Therefore, in this model, the σ_j^2 satisfy

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^n X_{ij}^2 \frac{\pi_j}{1-\pi_j},$$

and it is easily shown that they admit good data-driven estimators $\hat{\sigma}_j^2$, see Rosenbaum and Tsybakov (2010).

The construction of our new estimator is based on the following idea. We cannot use X in our estimation procedure since only its noisy version Z is available. In particular, the MU selector involves the matrix Z^TZ/n instead of X^TX/n. Compare to X^TX/n, this matrix contains a bias induced by the diagonal entries of the matrix $\Xi^T\Xi/n$ whose expectations σ_j^2 do not vanish. Therefore, if the σ_j^2 can be estimated, a natural idea is to compensate this bias thanks to these estimates. This leads to a new estimator $\hat{\theta}$ called **Compensated MU selector** and defined as a solution of the minimization problem

$$\min\{|\theta|_1 : \ \theta \in \Theta, \left|\frac{1}{n}Z^T(y - Z\theta) + \widehat{D}\theta\right|_\infty \le \mu|\theta|_1 + \tau\},$$

where $\widehat{D}$ is the diagonal matrix with entries $\hat{\sigma}_j^2$, which are estimators of σ_j^2, and $\mu \ge 0$ and $\tau \ge 0$ are constants chosen according to the level of the noises and the accuracy of the $\hat{\sigma}_j^2$.

In particular, this modification of the MU selector enables us to obtain bounds for the estimation errors which are decreasing with n. This is in contrast to the case of the MU selector, where the bounds are small only if the noise Ξ is small.

These estimation bounds for the Compensated MU selector are determined by the intensity of the noises, the accuracy of the $\hat{\sigma}_j^2$ and by the properties of the Gram matrix

$$\Psi = \frac{1}{n}X^TX.$$

For a vector θ, we denote by θ_J the vector in $\mathbb{R}^p$ that has the same coordinates as θ on the set of indices $J \subset \{1,\dots,p\}$ and zero coordinates on its complement J^c. We denote by $|J|$ the cardinality of J.

To state our results, following Gautier and Tsybakov (2011), we use the sensitivity characteristics related to the action of the matrix Ψ on the cone

$$C_J = \{\Delta \in \mathbb{R}^p : \ |\Delta_{J^c}|_1 \le |\Delta_J|_1\},$$

where J is a subset of $\{1,\dots,p\}$. For $q \in [1,\infty]$ and an integer $s \in [1,p]$, we define the ℓ_q *sensitivity* as follows:

$$\kappa_q(s) = \min_{J:\ |J|\le s}\Big(\min_{\Delta\in C_J:\ |\Delta|_q=1} |\Psi\Delta|_\infty\Big).$$

We will also consider the *coordinate-wise sensitivities*

$$\kappa_k^*(s) = \min_{J:\ |J|\le s}\Big(\min_{\Delta\in C_J:\ \Delta_k=1} |\Psi\Delta|_\infty\Big),$$

where Δ_k is the kth coordinate of Δ, $k = 1,\dots,p$. We will assume in the following the positivity of $\kappa_q(s)$ (or $\kappa_k^*(s)$). This requirement is weaker than the usual assumptions related to the structure of the Gram matrix Ψ, such as the Restricted Eigenvalue assumption and the Coherence assumption, see Gautier and Tsybakov (2011).

We have the following estimation bounds for the Compensated MU selector.

Theorem 1. *Assume $\theta^* \in \Theta$ is s-sparse. Then, with probability at least $1-\varepsilon$, we have*

$$\begin{aligned} |\hat{\theta}-\theta^*|_q &\leq \frac{\nu(\varepsilon)}{\kappa_q(s)}, \quad \forall\ 1 \leq q \leq \infty, \\ |\hat{\theta}_k-\theta^*_k| &\leq \frac{\nu(\varepsilon)}{\kappa^*_k(s)}, \quad \forall\ 1 \leq k \leq p, \\ \frac{1}{n}|X(\hat{\theta}-\theta^*)|_2^2 &\leq \min\big\{\frac{\nu^2(\varepsilon)}{\kappa_1(s)},\ 2\nu(\varepsilon)|\theta^*|_1\big\}. \end{aligned}$$

Moreover, in the model with missing data, if the components of ξ and Ξ are subgaussian, $\nu(\varepsilon)$ is of order $O(n^{-1/2})$ up to some logarithmic factor.

We also consider a variation of the Compensated MU selector. We set

$$W_0 = \{(\theta,u):\ |\frac{1}{n}Z^T(y-Z\theta)+\widehat{D}\theta+u|_\infty \leq c_1\tau\},$$

where $c_1 \geq 1$ is a suitably chosen constant. Then we consider (θ',u') a solution of the minimization problem

$$\min\{|\theta|_1 + \frac{1}{\lambda\sqrt{\mu}}|u|_\infty:\ (\theta,u) \in W_0\},$$

where λ is a tuning parameter. Eventually, our second estimator $\tilde{\theta}$ is defined such that $(\tilde{\theta},\tilde{u})$ is a solution of

$$\min\{|\theta|_1:\ (\theta,u) \in W_0,\ |u|_\infty \leq \big(\frac{|\theta'|_1+c_2\tau}{1-c_3\sqrt{\mu}}\big)\}.$$

Asymptotically, this estimator essentially shares the same properties as the Compensated MU selector.

References

[1] E. Gautier and A. B. Tsybakov, *High-dimensional instrumental variables regression and con- fidence sets*, arxiv:1105.2454

[2] M. Rosenbaum and A. B. Tsybakov, *Sparse Recovery under Matrix Uncertainty*, The Annals of Statistics **38** (2010), 2620–2651.

[3] M. Rosenbaum and A. B. Tsybakov, *Improved Matrix Uncertainty Selector*, In revision, IMS Collections, Festrschrift in honor of Jon Wellner.

Asymptotic behaviour of empirical Bayes procedures

Harry van Zanten
(joint work with Botond Szabó, Aad van der Vaart)

In recent years we have seen increasing use of Bayesian methods in high-dimensional or nonparametric statistical problems. It is by now well known that the (asymptotic) performance of such methods is sensitive to the fine properties of the priors that are used. Even seemingly reasonable choices may lead to inconsistent or suboptimal procedures. Priors used in Bayesian nonparametrics typically depend on one or more tuning parameters, so-called hyperparameters. In the case of function estimation such parameters can for instance describe the degree of regularity of a prior, a length scale, or a bandwidth. As a consequence of the general sensitivity of nonparametric Bayes procedures to the choice of the prior, hyperparameters must be tuned very carefully in order to ensure good performance of the resulting procedure.

Using a fixed value for a hyperparameter is generally undesirable, since it makes it likely that the prior is not properly matched to the true parameter of interest. Therefore practitioners typically favour data-driven choices. Two big classes of methods are widely used. The first is to endow the hyperparameters with a prior distribution as well. This leads to fully Bayesian procedures using so-called hierarchical priors. The frequentist behaviour of such methods has recently been studied in number of papers, and it was found that, if the priors are carefully constructed, they can yield adaptive, rate-optimal procedures for a range of nonparametric statistical problems. A second possible approach is to estimate the hyperparameters from the data, for instance using a likelihood-based method. This approach is not fully Bayesian, and commonly called *empirical Bayes*, but is often computationally convenient and therefore widely used in practice.

The theoretical performance of empirical Bayes methods in nonparametric problems has only been studied in a limited number of special cases. A general perspective on the frequentist behaviour of empirical Bayes methods has however been lacking until now. In the paper [1] we study the theoretical performance of empirical Bayes in the setting of the signal-in-white noise model and using a Gaussian prior involving a multiplicative scaling hyperparameter. We investigate in detail how the performance of the empirical Bayes method compares to an asymptotically optimal *oracle* procedure.

In general our results are favourable to the empirical Bayes method. However, the situation is delicate and we exhibit some surprising behaviour. For certain combinations of true parameters and (unscaled) priors the empirical Bayes approach yields a rate-optimal, adaptive procedure, whereas for other combinations it performs sub-optimally. In the language of function estimation we find that an empirical Bayes procedure works optimally if the unscaled prior oversmoothes or only slightly undersmoothes the unknown function of interest. If the initial prior is chosen too rough however, we obtain sub-optimal convergence rates. This shows that for empirical Bayes methods as well, great care needs to be taken when designing procedures for high-dimensional statistical problems.

References

[1] B. Szabó, A.W. van der Vaart and J.H. van Zanten. *Understanding the asymptotic behaviour of empirical Bayes procedures*, preprint.

Joint variable and rank selection for parsimonious estimation of high dimensional matrices

Marten H. Wegkamp

(joint work with Florentina Bunea, Yiyuan She)

This talk is devoted to optimal dimension reduction methods for sparse, high dimensional multivariate response regression models. Both the number of responses and that of the predictors may exceed the sample size. Sometimes viewed as complementary, predictor selection and rank reduction are the most popular strategies for obtaining lower dimensional approximations of the parameter matrix in such models. We show that important gains in prediction accuracy can be obtained by considering them jointly. For this, we first motivate a new class of sparse multivariate regression models, in which the coefficient matrix has both low rank and zero rows or can be well approximated by such a matrix. Then, we introduce estimators that are based on penalized least squares, with novel penalties that impose simultaneous row and rank restrictions on the coefficient matrix. We prove that these estimators indeed adapt to the unknown matrix sparsity and have fast rates of convergence. Our theoretical results are supported by a simulation study.

References

[1] F. Bunea, Y. She, M.H. Wegkamp, *Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices*, The Annals of Statistics **39**(2) (2011), 1282–1309.

[2] F. Bunea, Y. She, M.H. Wegkamp, *Joint Variable and Rank Selection for Parsimonious Estimation of High-Dimensional Matrices*, Preprint.

Statistical inference for high-dimensional data

Cun-Hui Zhang

(joint work with Tingni Sun, Stephanie Zhang)

We propose a semi low-dimensional (LD) approach for statistical analysis of certain types of high-dimensional (HD) data. The proposed approach is best described with the following model statement:

$$\text{model} = \text{LD component} + \text{HD component}. \tag{1}$$

The main objective of this semi-LD approach is to develop statistical inference procedures for the LD component, including p-values and confidence regions. This semi-LD approach is very much inspired by the semiparametric approach [1] in which a statistical model is decomposed as follows:

$$\text{model} = \text{parametric component} + \text{nonparametric component}.$$

Just as in the semiparametric approach, the worst LD submodel gives the minimum Fisher information for the LD component in (1), along with an efficient score function [6]. The efficient score function, or an estimate of it, can be used to derive an efficient estimator for the LD component. The efficient estimator is asymptotically normal with the inverse of the minimum Fisher information as its asymptotic covariance matrix. This asymptotic covariance matrix may be consistently estimated in a natural way. Consequently, approximate confidence intervals and p-values can be constructed based on the asymptotic theory.

Suppose we observe iid data with log-likelihood $\ell_i(\boldsymbol{\beta}) = \ell_i(\boldsymbol{\beta}|\text{data}_i)$ with an HD unknown $\boldsymbol{\beta}$. The Fisher information operator at $\boldsymbol{\beta}$ is

$$\boldsymbol{F} = -E_{\boldsymbol{\beta}}\ddot{\ell}_i(\boldsymbol{\beta}),\ \ddot{\ell}_i(\boldsymbol{\beta}) = (\partial/\partial\boldsymbol{\beta})(\partial/\partial\boldsymbol{\beta})^T\ell_i(\boldsymbol{\beta}).$$

For the estimation of a real parameter $\theta(\boldsymbol{\beta})$, consider one-dimensional submodels $\{\boldsymbol{\beta}+\boldsymbol{u}\phi, |\phi| < \epsilon_*\}$ with $\epsilon_* \to 0+$ slowly. Let $\boldsymbol{a}_0 = (\partial/\partial\boldsymbol{\beta})\theta(\boldsymbol{\beta})$. We impose the restriction $\boldsymbol{a}_0^T\boldsymbol{u} = 1$ so that $\theta(\boldsymbol{\beta}+\boldsymbol{u}\phi) - \theta(\boldsymbol{\beta}) \approx \phi$. The least favorable submodel at $\boldsymbol{\beta}$ is then given by $\boldsymbol{\beta}+\boldsymbol{u}_0\phi$ with

$$\boldsymbol{u}_0 = \arg\min_{\boldsymbol{u}}\big\{\boldsymbol{u}^T\boldsymbol{F}\boldsymbol{u} : \boldsymbol{a}_0^T\boldsymbol{u} = 1\big\} = \boldsymbol{F}^{-1}\boldsymbol{a}_0/(\boldsymbol{a}_0^T\boldsymbol{F}^{-1}\boldsymbol{a}_0).$$

The minimum Fisher information for the estimation of $\theta(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ is

$$F_0 = \boldsymbol{u}_0^T\boldsymbol{F}\boldsymbol{u}_0 = 1/(\boldsymbol{a}_0^T\boldsymbol{F}^{-1}\boldsymbol{a}_0).$$

Suppose we have the knowledge that $\boldsymbol{\beta} \approx \boldsymbol{b}_0$. We write the semi-LD model as

$$\boldsymbol{\beta} - \boldsymbol{b}_0 = \boldsymbol{u}_0\phi + \boldsymbol{\nu}, \quad \boldsymbol{a}_0^T\boldsymbol{\nu} = 0,$$

with unknown $\phi = \boldsymbol{a}_0^T(\boldsymbol{\beta}-\boldsymbol{b}_0)$ and $\boldsymbol{\nu} = (\boldsymbol{\beta}-\boldsymbol{b}_0) - \boldsymbol{u}_0\boldsymbol{a}_0^T(\boldsymbol{\beta}-\boldsymbol{b}_0)$. For small ϕ and $\boldsymbol{\nu}$, the maximum likelihood estimator (MLE) in the least favorable submodel is a natural candidate as an efficient estimator of ϕ, just as in linear regression with an orthogonal design. This leads to the following LD projection estimator (LDPE),

$$\widehat{\theta} = \theta(\widehat{\boldsymbol{\beta}}^{(init)}) + \arg\max_{\phi}\sum_{i=1}^n \ell_i\big(\widehat{\boldsymbol{\beta}}^{(init)} + \boldsymbol{u}\phi\big), \tag{2}$$

with suitable $\widehat{\boldsymbol{\beta}}^{(init)} \approx \boldsymbol{\beta}$ and a $\boldsymbol{u} \approx \boldsymbol{u}_0$. The LDPE (2) is efficient in the sense of

$$\sqrt{nF_0}(\widehat{\theta} - \theta) \to N(0,1), \tag{3}$$

provided an analysis of the log-likelihood in the local semi-LD model up to the order of $o_P(n^{1/2})$. This analysis has been carried out in linear regression in [7, 8].

In linear regression, we observe $(\boldsymbol{X},\boldsymbol{y}) \in \mathbb{R}^{(n+1)\times p}$ with $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{n\times n})$. The following scaled Lasso has been proposed in [7],

$$\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\} = \arg\min_{\boldsymbol{b},\sigma}\Big\{\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\|\boldsymbol{b}\|_1\Big\}, \tag{4}$$

with $\lambda_0 = 2\sqrt{(\log p)/n}$. The asymptotic efficiency of $\widehat{\sigma}$, $\widehat{\sigma}^2/\sigma^2 = \chi_n^2/n + o(n^{-1/2})$, has been proved there under an ℓ_2-regularity condition on $\boldsymbol{X}$ [9] and the following capped ℓ_1 sparsity condition on $\boldsymbol{\beta}$:

$$\sum_{j=1}^{p} \min\big(|\beta_j|/(\sigma\lambda_0), 1\big\} \le s$$

with $s \log p \ll n^{1/2}$. Under the same set of regularity conditions, methodologies for the statistical inference of regression coefficients has been developed in [8] based on the LDPE

$$\widehat{\beta}_j = \widehat{\beta}_j^{(init)} + \{\boldsymbol{z}_j^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)})\}/(\boldsymbol{z}_j^T\boldsymbol{x}_j). \tag{5}$$

This can be viewed as bias correction from the initial estimator. The success of the LDPE hinges on finding appropriate $\boldsymbol{z}_j$ with $\boldsymbol{z}_j^T\boldsymbol{X}/(\boldsymbol{z}_j^T\boldsymbol{x}_j) \approx \boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the unit vector in the j-th coordinate. We break the error as $\|\boldsymbol{z}_j^T\boldsymbol{X}/(\boldsymbol{z}_j^T\boldsymbol{x}_j) - \boldsymbol{e}_j\|_\infty\|\widehat{\boldsymbol{\beta}}^{(init)} - \boldsymbol{\beta}\|_1$ to take advantage of the assumed sparsity of $\boldsymbol{\beta}$. The scaled Lasso is used in [8] to find $\widehat{\boldsymbol{\beta}}^{(init)}$, $\widehat{\sigma}$ and $\boldsymbol{z}_j$. More recently, we proved the asymptotic efficiency of an LDPE $\widehat{\beta}_j$ in linear regression with random design under the weaker condition $s \log p \ll n$. In our analysis, all quantities are allowed to depend on n and $p \gg n \gg s \to \infty$ is permitted.

The above results essentially turn the regression problem into a much better understood Gaussian sequence problem with an $N(\boldsymbol{\beta}, (\sigma^2/n)\boldsymbol{V})$ observation, a known covariance structure $\boldsymbol{V} = (V_{jk})$, with $V_{jk} = n\boldsymbol{z}_j^T\boldsymbol{z}_k/\{\boldsymbol{z}_j^T\boldsymbol{x}_j)(\boldsymbol{z}_k^T\boldsymbol{x}_k)\}$, and a consistent estimator of the noise level σ.

The LDPE (5) is not sparse but can be used as a raw estimator for post processing for purposes depending on specific applications. Such post processing, possibly taking advantage of the sparsity of $\boldsymbol{\beta}$, will typically have a much clearer effect on each $\widehat{\beta}_j$, compared with existing methods. For example, approximate p-values for individual β_j based on the LDPE theory will still be valid after thresholding the raw $\widehat{\beta}_j$.

Among existing results, variable selection consistency is most relevant to statistical inference. We refer to recent reviews in [4, 9] and the recent book [2]. Consistent variable selection allows a great reduction of the complexity of the analysis from a large-p-smaller-n problem to one involving the oracle set of nonzero regression coefficients only. Consequently, taking the least squares estimator on the selected set of variables if necessary, statistical inference can be justified in the smaller oracle model. However, statistical inference based on selection consistency theory typically requires that all nonzero regression coefficients be greater than a noise level inflated to take model uncertainly into account. This assumption of uniform signal strength is, unfortunately, seldom supported by either the data or the underlying science, especially in biological and medical applications. In comparison, the capped ℓ_1 sparsity allows $\boldsymbol{\beta}$ to have many small elements.

The above semi-LD approach is parallel to the familiar semiparametric one [1]. In both cases, the main technical difficulty is to control the effect of the error of the initial estimator, possibly jointly with the effect of the error in the estimation of the efficient score [3, 5, 7, 8]. The main difference between a semi-LD analysis and existing semiparametric analyses is the lack of the knowledge of a manageable subspace for the approximation of the nuisance parameter. In a semiparametric analysis, the nonparametric component is typically well approximated by a known subspace of substantially smaller dimension than n. In a semi-LD analysis, the HD component is typically only known to be sparse, so that the noise is inflated by the uncertainty of not knowing the approximating subspace.

References

[1] P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, (1998), Springer, New York.

[2] P. Bühlmann and S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, (2011), Springer, New York.

[3] H. Chen, *Convergence rates for parametric components in a partly linear model*, The Annals of Statistics, **16** (1988), 136-146.

[4] J. Fan and J. Lv *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica, **20** (2010), 101-148.

[5] J.M. Robins and Y. Ritov, *Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models*, Statistics in medicine, **16** (1997), 285-319.

[6] C. Stein, *Efficient nonparametric testing and estimation*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, **1** (1956), 187-196.

[7] T. Sun and C.-H. Zhang, *Scaled Sparse Linear Regression*, arXiv, (2011), arXiv:1104.4595.

[8] C.-H. Zhang and S.S. Zhang, *Confidence intervals for low-dimensional parameters with high-dimensional data*, arXiv, (2011) arXiv:1110.2563.

[9] C.-H. Zhang and T. Zhang, *A general theory of concave regularization for high dimensional sparse estimation problems*, arXiv (2011) arXiv:1108.4988.

Model Selection and Sharp Asymptotic Minimaxity

Huibin Zhou

We study model selection for Gaussian Sequence model. The mean parameter is in a sparse l_p ball. Penalized procedures of the type $ck \log \frac{n}{k}$ are considered. We show $c = 2$ leads to sharp asymptotic minimaxity. As a consequence we proved a conjecture in Abramovich, Benjamini, Donoho and Johnstone (2006, Annals of Statistics).

References

[1] F. Abramovich, Y. Benjamini, D.L. Donoho and I.M. Johnstone *Adapting to unknown sparsity by controlling the false discovery rate*, The Annals of Statistics **34** (2006), 584-653.

Reporter: Markus Pauly

Participants

Prof. Dr. Yannick Baraud
Laboratoire J.-A. Dieudonne
Universite de Nice
Sophia Antipolis
Parc Valrose
F-06108 Nice Cedex 2

Markus Bibinger
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

Dr. Axel Bücher
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

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

Prof. Dr. T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia , PA 19104-6340
USA

Prof. Dr. Ismael Castillo
Laboratoire de Probabilites et
Modeles Aleatoires
Universite Paris VII
175 rue du Chevaleret
F-75013 Paris Cedex

Prof. Dr. Eustasio del Barrio Tellado
Departamento de Estadistica e
Investigacion Operativa
Facultad de Ciencias
C/Prado de la Magdalena s/n
E-47005 Valladolid

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Dr. Thorsten Dickhaus
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

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

Dr. Holger Drees
Department Mathematik
Universität Hamburg
Bundesstr. 55
20146 Hamburg

Prof. Dr. Lutz Dümbgen
Institut für mathematische Statistik &
Versicherungslehre
Universität Bern
Alpeneggstr. 22
CH-3012 Bern

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

Prof. Dr. Subhashis Ghoshal
Department of Statistics
North Carolina State University
Raleigh , NC 27695-8203
USA

Prof. Dr. Christophe Giraud
Centre de Mathematiques Appliquees
UMR 7641 - CNRS
Ecole Polytechnique
F-91128 Palaiseau Cedex

Dr. Shota Gugushvili
Department of Mathematics
Vrije University
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Philipp Heesen
Mathematisches Institut
Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
40225 Düsseldorf

Prof. Dr. Joel L. Horowitz
Northwestern University
Department of Economics
2001 Sheridan Road
Evanston , IL 60208-2600
USA

Prof. Dr. Arnold Janssen
Mathematisches Institut
Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
40225 Düsseldorf

Dr. Carsten Jentsch
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68131 Mannheim

Prof. Dr. Jiashun Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh , PA 15213
USA

Prof. Dr. Geurt Jongbloed
Delft Institute of Applied
Mathematics
Delft University of Technology
Mekelweg 4
NL-2628 CD Delft

Prof. Dr. Estate Khmaladze
School of Mathematics, Statistics and
Operations Research
Victoria University of Wellington
PO Box 600
Wellington 6140
NEW ZEALAND

Dr. Bas Kleijn
Korteweg-de Vries Instituut
Universiteit van Amsterdam
Postbus 94248
NL-1090 GE Amsterdam

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

Andreas Knoch
Mathematisches Institut
Heinrich-Heine-Universität
Gebäude 25.22
Universitätsstr. 1
40225 Düsseldorf

Prof. Dr. Arne Kovac
School of Mathematics
University of Bristol
University Walk
GB-Bristol BS8 1TW

Prof. Dr. Jens-Peter Kreiß
Institut für Mathematische
Stochastik der TU Braunschweig
Pockelsstr. 14
38106 Braunschweig

Prof. Dr. Mark Low
University of Pennsylvania
Department of Statistics
The Wharton School
Philadelphia PA 19104-6302
USA

Prof. Dr. Shuangge Ma
Division of Biostatistics
School of Public Health
Yale University
60 College St. LEPH 209
New Haven CT 06520
USA

Prof. Dr. Alexander Meister
Fachbereich Mathematik
Universität Rostock
18051 Rostock

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

Prof. Dr. Natalie Neumeyer
Department Mathematik
Universität Hamburg
Bundesstr. 55
20146 Hamburg

Prof. Dr. Andrew B. Nobel
Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill , NC 27599-3260
USA

Prof. Dr. Michael Nussbaum
Department of Mathematics
Cornell University
Malott Hall
Ithaca , NY 14853-4201
USA

Dr. Markus Pauly
Mathematisches Institut
Heinrich-Heine-Universität
Gebäude 25.22
Universitätsstr. 1
40225 Düsseldorf

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin

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

Prof. Dr. James M. Robins
Department of Biostatistics
Harvard School of Public Health
677 Huntington Ave.
Boston , MA 02115
USA

Prof. Dr. Judith Rousseau
Universite Paris Dauphine
Place du Marechal DeLattre de Tassigny
F-75016 Paris

Dr. Johannes Schmidt-Hieber
Department of Mathematics
Vrije University
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Dr. Dominic Schuhmacher
Institut für mathematische Statistik &
Versicherungslehre
Universität Bern
Alpeneggstr. 22
CH-3012 Bern

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

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

Prof. Dr. Aad W. van der Vaart
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
NL-2300 RA Leiden

Prof. Dr. Marten Wegkamp
Department of Mathematics
Cornell University
White Hall
Ithaca , NY 14853-7901
USA

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

Prof. Dr. Harry van Zanten
Department of Mathematics
Eindhoven University of Technology
P.O.Box 513
NL-5600 MB Eindhoven

Prof. Dr. Cun-Hui Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway , NJ 08854-8019
USA

Prof. Dr. Tong Zhang
Department of Statistics
Rutgers University
110 Frelinghuysen Road
Piscataway , NJ 08854-8019
USA

Mayya Zhilova
Weierstraß-Institut für
Angewandte Analysis und Stochastik
Mohrenstr. 39
10117 Berlin

Prof. Dr. Huibin Zhou
Department of Statistics
Yale University
P.O.Box 208290
New Haven , CT 06520-8290
USA