Mathematisches Forschungsinstitut Oberwolfach

Report No. 13/2014

DOI: 10.4171/OWR/2014/13

Adaptive Statistical Inference

Organised by
Mark Low, Philadelphia
Axel Munk, Göttingen
Alexandre Tsybakov, Paris

9 March – 15 March 2014

Abstract. This workshop in mathematical statistics highlights recent advances in adaptive methods for statistical estimation, testing and confidence sets. Related open mathematical problems are discussed with potential impact on the development of computationally efficient algorithms of data processing under prior uncertainty. Parcticular emphasis is on high dimensional models, inverse problems and discrtete structures.

Mathematics Subject Classification (2010): 62Gxx, 62Cxx, 62Fxx, 62Mxx,60Fxx, 60Gxx,

Introduction by the Organisers

The workshop *Adaptive Statistical Inference*, organised by Mark Low (Wharton), Axel Munk (Göttingen) and Alexandre Tsybakov (Paris) was attended by over 50 participants. The majority of the talks presented at this workshop can be clustered in the following thematic groups: adaptive nonparametric estimation in regression and machine learning, adaptation to the unkown sparsity in high-dimensional models, adaptive testing, adaptation to the unknown function in statistical inverse problems, adaptation for confidence and credible sets. Two larger survey talks have been given by O. Lepski on "Adaptive estimation over anisotropic functional classes via oracle approach" and by A. van der Vaart on "Confidence in Credible Sets". PhD students presented their work in a "Young researcher's session" Tuesday evening. On Wednesday a memorial session dedicated to the remembrance of Laurent Cavalier (Marseille) was held.

Adaptive nonparametric estimation.

The talk of Oleg Lepski deals with adaptive estimation in gaussian white noise model. It gives a classification of optimal rates of convergence of estimators on

anisotropic Besov and Nikolskii classes under L_q-norms. The method is based on new oracle inequalities for aggregation of linear estimators.

The talk of Lucien Birgé concentrates on adaptation to unknown distribution of errors in nonparametric regression. A universal procedure is suggested that adaptively achieves the optimal rates in the Helliger distance independently of the form of the noise.

Johannes Schmit-Hieber studies the problem of simultaneous adaptive estimation in L_2 and L_∞ norms, and he suggests a new procedure based on wavelet thresholding that achieves this task.

Alexander Rakhlin addresses the problem of comparison of the behavior of minimax risk in nonparametric estimation and minimax regret in statistical learning, showing that the rates for both quantities coincide when the complexities of the underlying functional classes are not too large.

Victor-Emmanuel Brunel gives a new insight into adaptive nonparametric estimation of convex sets and convex polytopes. He derives the optimal rates of convergence on classes of polytopes and suggests an adaptive procedure attaining this rate.

High-dimensional inference in regression models.

The talk of Arnak Dalalyan reports some refined results on the behavior of the Lasso estimators in high-dimensional linear regression when the Gram matrix of the problem is of low rank. In particular, the estimators are shown to automatically adapt to small rank in the sense of improving the prediction error.

The talk of Sara van de Geer is devoted to asymptotic confidence intervals for the parameters of high-dimensional linear regression based on Lasso type estimators.

Guillaume Lecué shows that the restricted eignevalue assumption, which is crucial for the study of Lasso-type methods, holds under weaker moment conditions than the sub-gaussianity supposed in the previous work.

Olga Klopp shows that varying coefficients models can be embedded into the framework of random matrix regression models. She proposes methods of estimation in this setting and derives optimal rates of convergence up to logarithmic factors.

Peter Bühlmann suggests an estimation procedure in high-dimensional mixture regression model. He introduces a summary parameter characterizing the mixture and studies the rates of estimation of this parameter when the dimension can be much larger than the sample size.

High-dimensional matrix models and graphical models.

The talk of Florentina Bunea deals with the problem of estimation of large covariance matrices having a banded structure. She introduces a new method, which is computationally feasible and adaptively attains the optimal rates both in Frobenius and operator norm.

Bin Yu discusses in her talk the problem of community detection in networks via a spectral clustering algorithm with regularisation. She shows adavantages

of the regularised algorithm as compared to the non-regularised one in terms of relaxing the assumptions on the model.

The talk of Anru Zhang is devoted to matrix completion in the situation when only a subset of rows and columns of an approximately low-rank matrix are observed. A new method of matrix recovery is proposed and it is shown that it achieves the optimal rate over certain classes of approximately low-rank matrices.

Harrison Zhou considers the problem of estimation of the individual entries of a precision matrix in Gaussian graphical model. He characterizes the conditions on the maximum degree of the graph, the dimension of the model and the sample size n such that each entry of the matrix can be adaptively estimated at the parametric $\sqrt{n}$ rate.

Adaptive testing hypotheses.

The talk of Michael Nussbaum discusses the methods of adaptive nonparametric testing on Sobolev ellipsoids of unknown radius.

Cristina Butucea considers testing hypotheses about large matrices observed in gaussian white noise. She derives minimax separation rates of testing the presence of a small cluster in such a matrix.

Statistical inverse problems.

Thorsten Hohage discusses Poisson inverse problems and highlights the contributions of Laurent Cavalier. He extended his work to nonlinear operators and analyzed nonlinear Tikhonov regularisation in this conetxt. To this end he showed an uniform exponential deviation inequality for the sup over a class of functionals of a Poisson proces.

Housen Li extended a multiscale estimation technique due to Nemirovski to deconvolution problems. He introduced a new estimator and a variational technique which allows to prove adaptation over a large scale of functions.

Tengyuan Liang studies the properties of the atomic norm constrained minimization procedure for the general statistical inverse problem setting. He also provides a lower bound on the minimax risk for this setting, which depends on dimension, sample size and volume ratio driven by the geometry of the model.

Adélaïde Olivier studies nonparametric estimation of the division rate function for a size-structured particle model from observing the life lengths of the particles that lived before a fixed time. She proposes a nonparametric estimator that attains the minimax optimal rate.

Markus Reiss analyzes the risk hull technique in an inhomogeneous sequence model. Estimation becomes first of all a model selection problem. He shows that the risk hull method gives adaptation to the unknown sparseness. Computational issues for the resulting estimators is addressed. This is based on previous joint work with Laurent Cavalier.

Confidence and Adaptation

Lutz Dümbgen revisits the classical problem of confidence bands for distribution functions. He constructs tighter confidence bands as the usual ones. This approach unifies the benefits of Kolmogoroff's and Berk-Jone's bands. To this end the

local uniform sub-excponential condition is established which allows to prove an exponential inequality of the sup of a calibrated process satisfying this condition.

Max Sommerfeld considers the problem of confidence statements for modes and their location in the context of circular data. He extends the SiZer methodology to this case and characetrizes the wrapped Gaussian kernel as the only one which generates a circular scale space. The connection to topological data analysis is highlighted.

Aad van der Vaart disucsses the question to what extend adaptation of credible sets can be achieved in the Bayesian nonparametric setting. He focuses on the empirical Bayes method to cope with unknown smoothness. This is analyszed in a sequence model which resembles polynomial degree of ill posedness. It is shown that a minimax contraction rate of the posterior in Soboloev classes is obtained. However, this has to fail when the smoothness index is unknown. Counterexamples for the validity of empirical Bayes are shown and weaker formulations of the problem are discussed. To this end the polished tail condition is introduced and a certain form of adaptation for empirical Bayes credible sets is shown.

Miscellaneous topics.

Robert Nowak studies the recovery of the best arm in multi-armed bandit problems. This work suggests a rate optimal algorithm, which is a modification of the upper confidence bound procedure using a finite sample version of the law of the iterated logarithm.

Laszlo Györfi closes the workshop by discussing an open problem for stationary gaussian time series. "Is it possible to learn the best predictor almost surely in a strongly consistent way?" Partial answers are given.

Workshop: Adaptive Statistical Inference

Table of Contents

Abstracts

A robust adaptive estimator for regression

Lucien Birgé
(joint work with Yannick Baraud and Mathieu Sart)

Our purpose is to present a new method for adaptively estimating a regression function when little is known about the shape and scale of the errors and that can cope with error distributions as different as Gaussian, Uniform, Cauchy or even with a unimodal unbounded density. To be more precise, let us describe the framework we want to deal with.

We observe n independent random variables $X_1, \dots, X_n$ each X_i with an unknown distribution P_i on a measurable space $(\mathscr{X}, \mathscr{A})$ and our aim is to use the vector $\mathbf{X} = (X_1, \dots, X_n)$ of observations to estimate their joint distribution $\mathbf{P} = \bigotimes_{i=1}^n P_i$, that is to find a random approximation $\widehat{\mathbf{P}}(\mathbf{X}) = \bigotimes_{i=1}^n \widehat{P}_i(\mathbf{X})$ of $\mathbf{P}$ based on the observed variables X_i. To measure the quality of the approximation of $\mathbf{P}$ by $\widehat{\mathbf{P}}$ we need a distance on the set of product measures on $\mathscr{X}^n$. It is known from Le Cam's work that a very convenient one is that (here denoted by $\mathbf{h}$) derived from the Hellinger distance h:

$$\mathbf{h}^2\left(\bigotimes_{i=1}^n P_i, \bigotimes_{i=1}^n Q_i\right) = \sum_{i=1}^n h^2(P_i, Q_i) = \frac{1}{2}\sum_{i=1}^n \int \left(\sqrt{dP_i} - \sqrt{dQ_i}\right)^2.$$

We recall that the Hellinger distance h is the bounded distance on the set of all probabilities on $\mathscr{X}$ given by

$$h^2(R,T) = \frac{1}{2}\int \left(\sqrt{\frac{dR}{d\mu}} - \sqrt{\frac{dT}{d\mu}}\right)^2 d\mu \le 1,$$

where μ is an arbitrary positive measure which dominates both R and T, the result being independent of the choice of μ.

We measure the quality of an estimator $\widehat{\mathbf{P}}$ by its *quadratic risk* $\mathbb{E}_{\mathbf{P}}[\mathbf{h}^2(\widehat{\mathbf{P}}(\mathbf{X}), \mathbf{P})]$ with respect to the distance $\mathbf{h}$, the notation $\mathbb{E}_{\mathbf{P}}$ meaning that $\mathbf{X}$ has the joint distribution $\mathbf{P}$. This framework is suitable for the analysis of *fixed design regression models* on $\mathbb{R}^n$, for which $\mathscr{X} = \mathbb{R}$ with $X_i = f_i + \varepsilon_i$ for $1 \le i \le n$, the ε_i being assumed to be i.i.d. with density p with respect to the Lebesgue measure μ. Therefore X_i has density $p(\cdot - f_i)$ with respect to μ and $\mathbf{X}$ the density $\mathbf{s} = \otimes_{i=1}^n p(\cdot - f_i)$ with respect to $\mu^{\otimes n}$.

The simplest case of fixed design regression occurs when the function $i \mapsto f_i$ is constant and equal to $\theta \in \Theta \subset \mathbb{R}$. It corresponds to the case of i.i.d. variables X_i with density $p(\cdot - \theta)$ and to a parametric model with a single translation parameter θ. The problem is then to find an estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ for θ so that

$$\widehat{\mathbf{P}} = P_{\widehat{\theta}}^{\otimes n} \quad \text{with} \quad \frac{dP_{\widehat{\theta}}}{d\mu} = p(\cdot - \widehat{\theta}).$$

If θ_0 obtains, that is $\mathbf{P} = P_{\theta_0}^{\otimes n}$, the quadratic risk of $\widehat{\mathbf{P}}$ or, equivalently, of $\widehat{\theta}$, writes

$$\mathbb{E}_{\theta_0}\left[\mathbf{h}^2\left(P_{\widehat{\theta}}^{\otimes n}, P_{\theta_0}^{\otimes n}\right)\right] = \mathbb{E}_{\theta_0}\left[nh^2\left(P_{\widehat{\theta}}, P_{\theta_0}\right)\right],$$

where $\mathbb{E}_\theta$ stands for $\mathbb{E}_{P_\theta^{\otimes n}}$. Since the risk is a function of the unknown parameter θ_0, a common way (although not the only one) of evaluating the performance of an estimator $\widehat{\theta}$ is via its maximum quadratic risk $R_M(\widehat{\theta}) = \sup_{\theta\in\Theta} \mathbb{E}_\theta[nh^2\left(P_{\widehat{\theta}}, P_\theta\right)]$. This leads to the notion of minimax risk for the problem at hand:

$$R_M(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta\left[nh^2\left(P_{\widehat{\theta}}, P_\theta\right)\right],$$

where the infimum runs over all possible estimators $\widehat{\theta}$ of θ. An optimal estimator $\widetilde{\theta}$ is therefore one that minimizes $R_M(\widehat{\theta})$. Unfortunately, computing $R_M(\Theta)$ exactly is generally an intractable optimization problem and we merely look for approximately optimal estimators $\widetilde{\theta}$ satisfying

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\left[nh^2\left(P_{\widetilde{\theta}}, P_\theta\right)\right] \leq CR_M(\Theta),$$

where C is a constant which does not depend on n.

Typically, classical estimators based on empirical moments or quantiles or the Maximum Likelihood Estimator do satisfy such requirements only in special cases (depending on the properties of p) and not the same for all estimators. Our new construction provides a rather general solution to this problem for all unimodal densities p when p is known but it can also deal with the case when p is only approximately known. It is based on a family of *models* $\overline{S}$ which are approximating sets for the unknown parameter f and it tends to select the best model (the one providing the best compromise between the approximation error of f by the model and the estimation error on the model, which usually depends on its size) among all of them. This construction is based on an estimation of the differences $\mathbf{h}^2(\mathbf{t}, \mathbf{s}) - \mathbf{h}^2(\mathbf{t}', \mathbf{s})$ for all points $\mathbf{t}, \mathbf{t}'$ belonging to the union of all available models. The relevant estimator for this quantity has been designed by Baraud in [1].

The complete construction and the analysis of its performance are to be found in arXiv : `http://arxiv.org/abs/1403.6057`. The procedure also applies to other statistical frameworks like density estimation and regression with a random design.

References

[1] Y. Baraud, *Estimator selection with respect to Hellinger-type risks*, Probab. Theory Relat. Fields **151** (2011), 353–401.

Adaptive estimation of convex polytopes and convex bodies

Victor-Emmanuel Brunel

We are interested in two models. The first one consists of observing a sample of i.i.d. random variables, with uniform distribution on some unknown subset of $\mathbb{R}^d, d \geq 1$. The second one consists of a regression setup, where the regression function is the indicator on some unknown subset of $\mathbb{R}^d, d \geq 1$.

In both cases, we estimate the unknown set under two possible assumptions. First, we assume that the unknown set is a convex polytope with r vertices, and $r \geq d+1$ is a known integer. Second, we assume that the unknown set is any convex body, and we give the corresponding minimax rate of convergence. In the polytopal case, if r is not known, we propose an adaptive estimator which achieves the same rate of convergence as in the known r case. In addition, we show that this adaptive estimator achieves the optimal rate of convergence in case of misspecification, i.e., when the true set is not a polytope, but any convex body. To finish, we discuss the optimality, in terms of rate of convergence, of our estimators in the polytopal and known r case.

References

[1] V.-E. Brunel, *Adaptive Estimation of Convex Polytopes and Convex Sets from Noisy Data*, Electronic Journal of Statistics **7** (2013), 1301–1327.
[2] V.E. Brunel, *Adaptive estimation of polytopal and convex support*, Submitted. (arXiv:1309.6602)
[3] V.E. Brunel, *A universal deviation inequality for random polytopes*, Submitted. (arXiv:1311.2902)
[4] A. P. Korostelev, A? B. Tsybakov, *Minimax Theory of Image Reconstruction*, Springer, NY (1993).

A new approach for large-scale inhomogeneous data

Peter Bühlmann
(joint work with Nicolai Meinshausen)

Our goal is to construct an estimator which can deal with inhomogeneities in large-scale data where the sample size and the dimension might be very large. Besides the challenge of dealing with heterogeneous data, we aim for a procedure which is computationally feasible for large scales.

We consider a mixture of regressions model:

$$Y_i = X_i^T B_i + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$, and the regression coefficient vector $B_i \in \mathbb{R}^p$ is allowed to change for every subject $i = 1, \ldots, n$. The B_i's are random variables from a distribution F_B. We always assume that the ε_i's are uncorrelated from the X_i's and the X_i's are independent from the B_i's so that there is no information from X to the mixture regression coefficients B. The following examples fit to the framework.

Known groups of observations. Every sample corresponds to a group $g \subseteq \{1, \ldots, n\}$, and there are G such groups which build a partition of $\{1, \ldots, n\}$. In every group g, we assume that $B_i \equiv b_g$ for all $i \in g$. Thus, the model in (1) is a finite mixture of G regression coefficients with known mixture components.

Correlated B_i's. When having strong positive correlation among neighboring B_i's (neighboring w.r.t. the sample ordering i), we have a smooth trend for the regression coefficients B_i's (w.r.t. i). The model in (1) is then a random coefficient linear model with smooth trend.

Contaminated samples. We might assume that a large fraction $(1 - \delta)$ of the B_i's assumes a fixed value b, and there is a smaller fraction δ of outliers where B_i's can take other values in $\mathbb{R}^p$. The model (1) can then be viewed from the viewpoint of robust statistics for guarding against some outliers, and the (groups of) outliers are unknown.

In [1] the following main issues are covered: (i) a definition of a new "maximin" parameter b_{maximin} which is an important "summary quantity" of all the possible values B_i so that prediction and interpretation in heterogeneous models remains powerful; (ii) establishing estimation rates for the "maximin" parameter b_{maximin}, covering also the high-dimensional setting where the dimension p might be much larger than sample size n; (iii) showing that the estimator can be computed, in some circumstances, with a very efficient linear program allowing for very large scales.

References

[1] N. Meinshausen and P. Bühlmann, *Maximin effects in inhomogeneous large-scale data*, Preprint.

Convex banding of the covriance matrix

Florentina Bunea

(joint work with Jacob Bien and Luo Xiao)

The estimation of large covariance matrices of a random vector with entries that are or can be ordered is an intensely studied problem in stochastic processes, spatial statistics and general high dimensional inference. When the population matrix is banded or approximately banded, a number of theoretically and practically optimal estimators have been proposed in the last six years. With very few exceptions, the theoretically optimal estimators are not adaptive, and the estimates that are practically performant do not have established theoretical properties. Moreover, most existing theoretical analyses are restricted to population covariance matrices with bounded operator norm.

We introduce a new estimator, the hierarchically banded estimator, which is the solution of a computationally feasible convex optimization problem. During this procedure, one successively penalizes nested triangular corners of the current candidate estimate, using the sample covariance matrix in the first step. Since these ensembles are nested, the new penalty is not a simple variation on the existing

group-type penalties and poses new challenges. We show that the procedure can be implemented successfully and efficiently. The proposed estimator achieves, adaptively, the minimax optimal convergence rates in Frobenius and operator norm. These results are established over classes of banded or semi-banded population matrices, and members of these classes are allowed to have diverging operator norm.

Formally, we assume to have observed $X_1, \ldots, X_n$ independent copies of a p-dimensional vector X with mean zero and covariance matrix Σ. We assume that the marginals of X have sub-Gaussian distribution. We let $\widehat{\Sigma}$ denote the sample covariance matrix. For a given tuning parameter λ, we define our estimator as

$$\hat{P} = \arg\min_{P} \left\{ \|P - \hat{\Sigma}\|_F^2 + \lambda \|P\|_{2,1}^* \right\}, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, and

$$\|P\|_{2,1}^* = \sum_{\ell=1}^{p-1} \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|P_{s_m}\|_2^2},$$

with

$$w_{\ell,m} = \frac{\sqrt{\ell}}{\ell - m + 1}, \text{ for } 1 \le m \le \ell,\ 1 \le \ell \le p-1,$$

and $s_m = \{(j,k) : \ |j-k| = p-m\}$. In [1] we showed that, although the original convex problem (1) is not separable, its dual is. We therefore employ a Block Coordinate Descent algorithm to solve the dual, and use the primal-dual relationship to reconstruct the solution to (1).

The following theorem shows that our estimator can also be regarded as a tapering estimator. However, the tapering function is not pre-specified in functional form prior to estimation, as in existing work, and is implicitly and recursively defined as below, in a fully data-dependent manner.

Theorem 1. *The convex banding estimator, $\hat{P}$, can be written as a tapering estimator with a Toeplitz, data-dependent tapering matrix, $\hat{P} = \hat{T} * \hat{\Sigma}$:*

$$\hat{T}_{s_m} = \begin{cases} 1_m & \text{for } m = p \text{ (diagonal)} \\ \prod_{\ell=m}^{p-1} \frac{[\hat{\nu}_\ell]_+}{w_{\ell m}^2 + [\hat{\nu}_\ell]_+} 1_m & \text{for } 1 \le m \le p-1 \end{cases}$$

where $\hat{\nu}_\ell$ satisfies $\lambda^2 = \sum_{m=1}^{\ell} \frac{w_{\ell m}^2}{(w_{\ell m}^2 + \hat{\nu}_\ell)^2} \|\hat{R}_{s_m}^{(\ell)}\|^2$, $\hat{R}^{(1)} = \hat{\Sigma}$ and for $\ell = 1, \ldots, p-2$, and for each $m \le \ell$, we have

$$\hat{R}_{s_m}^{(\ell+1)} \quad = \quad \frac{[\hat{\nu}_\ell]_+}{w_{\ell m}^2 + [\hat{\nu}_\ell]_+} \hat{R}_{s_m}^{(\ell)}, \tag{2}$$

and $1_m \in R^m$ denotes the vector of ones.

Immediate consequences of this theorem are: (1) The estimator $\hat{P}$ is, as desired, banded; (2) $\hat{P}$ is positive definite, with high probability.

We showed in [1] that, if the population covariance matrix has bandwidth K, this value will be recovered by the bandwidth of the estimator, with high probability, under minimal signal strength conditions on the entries of Σ.

The following theorem is an oracle inequality, with leading constant one for the bias term. It shows that the proposed estimator achieves the best bias-variance trade-off with respect to the Frobenius norm, with high probability.

Theorem 2. *If* $\max_{i,j} |\Sigma_{ij}| \leq M$*, for some constant* M *and* $\lambda = x\sqrt{\log p/n}$ *then, with high probability,*

$$\|\hat{P} - \Sigma\|_F^2 \leq \inf_{B \in \mathbb{R}^{p \times p}} \left\{ \|\Sigma - B\|_F^2 + C_1 \frac{K(B) p \log p}{n} \right\} + x^2 \frac{p \log p}{n},$$

for some constant C_1*, and where* $K(B)$ *denotes the bandwidth of a generic matrix* B*.*

From this theorem one obtains immediately the minimax optimal rates, up to logarithmic terms, over the class of exactly banded matrices and over the class of approximately banded matrices:

$$\frac{\|\hat{P} - \Sigma\|_F^2}{p} \leq CK \log(p)/n,$$

and

$$\frac{\|\hat{P} - \Sigma\|_F^2}{p} \leq C \left(\frac{\log p}{n} \right)^{\frac{2\alpha+1}{2\alpha+2}}.$$

Moreover, Theorem 2 shows that this rate analysis can be performed directly over the unifying framework provided by the class of semi-banded matrices, which we introduced in [1] and give below:

$$\mathcal{G}(K) = \left\{ \Sigma : \max_{ij} |\Sigma_{ij}| \leq M \text{ and } \|\Sigma - B\|_F^2 \leq CpK \log p/n, \right\} \tag{3}$$

for some banded B with bandwidth K, and some constant $C > 0$.

We also show that, over the class of banded matrices with possibly diverging bandwidth K, our estimator achieves adaptively the minimax rate in operator norm, up to logarithmic factors:

$$\|\hat{P} - \Sigma\|_{op} \leq CK\sqrt{\log p/n},$$

with high probability, under a minimal signal strength condition.

References

[1] J. Bien, F. Bunea and L. Xiao, *Convex banding of the covariance matrix*, on arXiv, (2014).

[2] L. Xiao, F. Bunea *On the theoretical and practical merits of the banding estimator for large covariance matrices*, on arXiv (2014).

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

Tests for high-dimensional sparse matrices

Cristina Butucea
(joint work with Yuri I. Ingster, Ghislaine Gayraud)

The talk is based on the papers [1] and [2], concerning sharp minimax tests in high-dimensional matrices containing a sparse signal structured as a submatrix.

In [1], we have a $N \times M$ matrix

$$Y_j = s_{ij} + \xi_{ij},\ i = 1, ..., M,\ j = 1, ..., N,$$

with ξ_{ij} i.i.d. with standard Gaussian distribution, and $s_{ij} \in \mathbb{R}$. We test the null hypothesis that there is no signal in the large matrix against the alternative that there exists some submatrix of size $n \times m$ with significant elements in the sense that $s_{ij} = a > 0$. We propose a test procedure and compute the asymptotical detection boundary a so that the maximal testing risk tends to 0 as m, $M \to \infty$, n, $N \to \infty$, $p = n/N \to 0$ and $q = m/M \to 0$. We prove that this boundary is sharp minimax asymptotically under some additional constraints. Relations with other testing problems are discussed. We propose a testing procedure which adapts to unknown (n, m) within some given set and compute the adaptive sharp rates.

In [2], we generalize the previous results to matrix-valued Gaussian sequence model, that is, we observe a sequence of high-dimensional $M \times N$ matrices of heterogeneous Gaussian random variables $x_{ij,k}$ for $i \in \{1, ..., M\}$, $j \in \{1, ..., N\}$ and $k \in \mathbb{Z}$. The standard deviation of our observations is ϵk^s for some $\epsilon > 0$ and $s \geq 0$. This model can be seen as a Gaussian white noise model where the signal is an additive function of $M \times N$ coordinates and observed in a mildly ill-posed inverse problem.

We give sharp rates for the detection of a sparse submatrix of size $m \times n$ with active components. A component (i, j) is said active if the sequence $\{x_{ij,k}\}_k$ has mean $\{\theta_{ij,k}\}_k$ within a Sobolev ellipsoid of smoothness $\tau > 0$ and total energy $\sum_k \theta^2_{ij,k}$ larger than some r^2_ϵ. Our rates involve relationships between m, n, M and N tending to infinity such that m/M, n/N and ϵ tend to 0, such that a test procedure that we construct has asymptotic minimax risk tending to 0.

We prove corresponding lower bounds under additional assumptions on the relative size of the submatrix in the large matrix of observations. Except for these additional conditions our rates are asymptotically sharp. Lower bounds for hypothesis testing problems mean that no test procedure can distinguish between the null hypothesis (no signal) and the alternative, i.e. the minimax risk for testing tends to 1.

References

[1] C. Butucea and Yu.I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix*, Bernoulli **19** (2013), 2652–2688.

[2] C. Butucea and G. Gayraud, *Sharp detection of smooth signals in a high-dimensional sparse matrix with indirect observations*, (2013), arxiv:1301.4660

On the Prediction Loss of the Lasso and Total Variation Penalization

Arnak S. Dalalyan
(joint work with Mohamed Hebiri and Johannes Lederer)

In recent years, considerable effort has been devoted to establishing sharp theoretical guarantees for the prediction performance of the Lasso. Although for a variety of settings various types of risk bounds are already available, the prediction performance of the Lasso is still not completely understood. In this work, we improve the sharpest known risk bounds to gain new insight into the prediction performance of the Lasso.

We study the prediction performance of the Lasso only for Gaussian linear regression models with deterministic design. More specifically, the model reads as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \sigma^* \mathcal{N}_n(0, \mathbf{I}_n), \tag{1}$$

where $\boldsymbol{y} := (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ is the response vector, $\mathbf{X} := (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ the design matrix (for which we assume, without loss of generality, that $\|\boldsymbol{x}^j\|_2^2 \le n$ for all $j \in \{1, \dots, p\}$), $\xi \in \mathbb{R}^n$ the noise vector, and $\mathbf{I}_n$ denotes the identity matrix. We recall that the Lasso is any solution of the convex optimization problem

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} \in \arg\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\}, \tag{2}$$

that can be efficiently solved even for very large values of p and n. The magnitude of the tuning parameter $\lambda > 0$ determines the amount of penalization and, therefore, has a crucial influence on the performance of the Lasso.

The main findings of the present work (see [1] for more details) can be summarized as follows.

(1) We prove that the Lasso estimator used with the universal choice of the tuning parameter $\lambda = \sqrt{2\log(p)/n}$ has a prediction loss at least proportional to $\frac{\log(p)}{n} \times \text{rank}(\mathbf{X})$, where $\text{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$.
(2) For sparse vectors $\boldsymbol{\beta}^*$ with support $J^* = \{j \in \{1, \dots, p\} : \beta_j^* \neq 0\}$ and for covariates that are strongly correlated in the sense that all irrelevant covariates $\{\boldsymbol{x}^j : j \notin J^*\}$ are close to the linear span of relevant covariates $\{\boldsymbol{x}^j : j \in J^*\}$, we show that choosing a tuning parameter λ that is substantially smaller than the universal one leads to a considerable gains in terms of prediction loss. We present a simple manner to incorporate the geometry of the covariates into the tuning parameter that provides fast rates of prediction when the covariates are strongly correlated.
(3) For really sparse vectors, that is, for s^* considerably smaller than n (for example, s^* is fixed and $n \to \infty$), there are methods that satisfy fast rate bounds for prediction irrespective of the correlations of the covariates. For Lasso prediction, we exhibit a counter-example showing that it is impossible to get fast rate bounds without imposing some relatively strong constraints on the correlations of the covariates. This is true even if we allow for oracle choices of the tuning parameter λ, that is, if we allow for λ that depend on the true regression vector $\boldsymbol{\beta}^*$ and the noise level σ^*.

(4) Finally, previously known results imply fast rates for prediction with the Lasso in the following two extreme cases: First, when the covariates are mutually orthogonal, and second, when the covariates are all collinear. But how far from these two extreme cases can a design be such that it still permits fast rates for prediction with the Lasso? For the first case, the case of mutually orthogonal covariates, this question has been thoroughly studied in the literature, whereas there were only a few results in the second case. We fill this gap by proving that if the irrelevant covariates are within a constant (Euclidean) distance of the linear span of the relevant covariates, then the Lasso atteins the fast rates of prediction.

As a consequence of the obtained risk bounds, we show that the total-variation penalized least-squares estimator achieves the nearly parametric rate $(\log n)^2/n$ when the unknown signal is piecewise constant.

References

[1] A. S. Dalalyan, M. Heiri, J. Lederer, *On the Prediction Performance of the Lasso*, arXiv, `http://arxiv.org/pdf/1402.1700v1.pdf` (2014).

Worst possible sub-directions in high-dimensional statistics

Sara van de Geer

This work is motivated by the need for confidence intervals in high-dimensional models. For the standard linear model a de-sparsifying technique has been introduced in [4] This technique has been extended to generalized linear models in [2]. We examine further extensions. As an example, consider an $n \times p$ data matrix X consisting of i.i.d. rows with distribution P. Let $\hat{\Sigma}$ be the empirical inner product matrix, and $\Sigma_0 := \mathbb{E}\hat{\Sigma}$ be the theoretical inner product matrix. We are interested in estimating the precision matrix $\Theta_0 := \Sigma_0^{-1}$. Let $\hat{\Theta}$ be an estimator, for example based on the graphical Lasso or the node-wise Lasso. Then as de-sparsified version, we propose

$$\hat{\Theta}_{\text{de-sparsified}} = \hat{\Theta} + \hat{\Theta}^T - \hat{\Theta}^T \hat{\Sigma} \hat{\Theta}.$$

One can now decompose:

$$\hat{\Theta}_{\text{de-sparsified}} - \Theta_0 = -\Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 - \text{rem}_1 - \text{rem}_2$$

where

$$\text{rem}_1 := (\hat{\Theta} - \Theta_0)^T(\hat{\Sigma} - \hat{\Theta}^{-1})\hat{\Theta} = (\hat{\Theta} - \Theta_0)^T(\hat{\Sigma}\hat{\Theta} - I)$$

and

$$\text{rem}_2 := \Theta_0(\hat{\Sigma} - \Sigma_0)(\hat{\Theta} - \hat{\Theta}_0).$$

The first term is linear in $\hat{\Sigma} - \Sigma_0$ and deriving asymptotic normality for each of its entries is straightforward. The challenge is now to prove that the two remainder terms can be neglected. For the graphical Lasso, this problem is studied in [1].

More generally, the main issue when studying single components or low-dimensional sub-components of the unknown parameter, is the anti-projection of the

information for the parameter of interest on the nuisance parameters. To clarify this in an example, consider the linear model

$$Y = X\beta^0 + \epsilon$$

and the Lasso

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta\|_1 \right\},$$

where $\|v\|_n^2 := v^T v/n$. The de-sparsified Lasso is

$$\hat{\beta}_{\text{de-sparsified}} = \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n$$

where $\hat{\Theta}$ is obtained by the node-wise Lasso. One now has the decomposition

$$\hat{\beta}_{\text{de-sparsified}} - \beta^0 = \hat{\Theta}^T X^T \epsilon/n - \text{rem}_1$$

where

$$\text{rem}_1 = (\hat{\Theta}^T \hat{\Sigma} - I)(\hat{\beta} - \beta^0).$$

This remainder term can be handled using the KKT-conditions. Let now, for $j \in \{1, \ldots, p\}$, X_j be the j-th column of X and $X_{-j} := \{X_k\}_{k \neq j}$ be the remaining columns. Let $X_j \hat{\text{P}} X_{-j} := X_{-j}\hat{\gamma}_j$ be the approximate projection of X_j on X_{-j} obtained using the Lasso for the regression of X_j on X_{-j}. Let $X_j \hat{\text{A}} X_{-j} := X_j - X_j \hat{\text{P}} X_{-j} = X\hat{c}_j$ be the approximate anti-projection. We call $\hat{c}_j$ the approximate worst possible sub direction for estimating β_j^0. The de-sparsified estimator is

$$\hat{\beta}_{j,\text{de-sparsified}} = \frac{(X_j \hat{\text{A}} X_{-j})^T (Y - X_{-j}\hat{\beta}_{-j})}{(X_j \hat{\text{A}} X_{-j})^T X_j}$$

where $\hat{\beta}_{-j} := \{\hat{\beta}_k\}_{k \neq j}$.

We show in [3] that

$$|\hat{\beta}_j - \beta_j^0| \leq \frac{|(X_j \hat{\text{A}} X_{-j})^T \epsilon|}{(X_j \hat{\text{A}} X_{-j})^T X_j} + \frac{\lambda \|\hat{\gamma}_j\|_1}{(X_j \hat{\text{A}} X_{-j})^T X_j/n} + \text{rem}_1$$

whereas

$$|\hat{\beta}_{j,\text{de-sparsified}} - \beta_j^0| \leq \frac{|(X_j \hat{\text{A}} X_{-j})^T \epsilon|}{(X_j \hat{\text{A}} X_{-j})^T X_j} + \text{rem}_1.$$

In other words, by de-sparsifying one removes the bias

$$\frac{\lambda \|\hat{\gamma}_j\|_1}{(X_j \hat{\text{A}} X_{-j})^T X_j/n}$$

due to the ℓ_1-penalty on coefficients of the approximate projection of X_j on the other variables. The remainder term rem_1 is moreover small due to the approximate orthogonality of $X_j \hat{A} X_{-j}$ and $X_j \hat{P} X_{-j}$.

Consider now a general high-dimensional model. We define worst possible sub-directions in a similar way and bounds for parameters of interest of the ℓ_1-penalized M-estimator that involve the ℓ_1-norm of the worst possible sub-direction. This serves as first step toward de-sparsifying which aims in a first stage at removing

this bias and in the second stage at showing that a linear, asymptotically normal term is dominating the other (remainder) term. As intermediate goal, the de-sparsified estimator can be used for variable selection without imposing irrepresentable conditions.

References

[1] J. Jankova and S. van de Geer, *Confidence intervals for high-dimensional inverse covariance estimation*, **http://arxiv.org/abs/1403.6752** (2014).

[2] S. van de Geer, P. Bühlmann, Y. Ritov and R. Dezeure, *On asymptotically optimal confidence regions and tests forf high-dimensional models*, The Annals of Statistics (2014), to appear.

[3] S. van de Geer, *Worst possible sub-directions in high-dimensional models*, **http://arxiv.org/abs/1403.7023** (2014).

[4] C.-H. Zhang and S.S. Zhang, *Confidence intervals for low-dimensional parameters in high-dimensional linear models*, Journal of the Royal Statistical Society Series B **76** (2014), 217–242.

Confidence Bands for a Distribution Function: A New Look at the Law of the Iterated Logarithm

Lutz Dümbgen

(joint work with Jon A. Wellner)

Let $\hat{F}_n$ be the empirical distribution function of independent random variables $X_1, X_2, \ldots, X_n$ with unknown distribution function F on the real line. It is well-known that the stochastic process $\big(\hat{F}_n(x)\big)_{x\in\mathbb{R}}$ has the same distribution as $\big(\hat{G}_n(F(x))\big)_{x\in\mathbb{R}}$, where $\hat{G}_n$ is the empirical distribution of independent random variables $U_1, U_2, \ldots, U_n$ with uniform distribution on $[0,1]$. This enables us to construct confidence bands for the distribution function F. A well-known classical method are Kolmogorov-Smirnov confidence bands: Let

$$\mathbb{U}_n(t) \;:=\; n^{1/2}(\hat{G}_n(t) - t),$$

and let $\kappa^{\mathrm{KS}}_{n,\alpha}$ be the $(1-\alpha)$-quantile of

$$\|\mathbb{U}_n\|_\infty \;:=\; \sup_{t\in[0,1]} |\mathbb{U}_n(t)|.$$

Then with probability at least $1-\alpha$,

$$F(x) \;\in\; [\hat{F}_n(x) \pm n^{-1/2}\kappa^{\mathrm{KS}}_{n,\alpha}] \quad \text{for all } x \in \mathbb{R}.$$

Equality holds if F is continuous. Since $\mathbb{U}_n$ converges in distribution in $\ell_\infty([0,1])$ to standard Brownian bridge $\mathbb{U}$, $\kappa^{\mathrm{KS}}_{n,\alpha}$ converges to the $(1-\alpha)$-quantile $\kappa^{\mathrm{KS}}_\alpha$ of $\|\mathbb{U}\|_\infty$. In particular, these simultaneous confidence intervals have width $O(n^{-1/2})$ uniformly in $x \in \mathbb{R}$.

Another method, based on a goodness-of-fit test by Berk and Jones (1979), was introduced by Owen (1995): Let $\kappa^{\mathrm{BJ}}_{n,\alpha}$ be the $(1-\alpha)$-quantile of

$$T^{\mathrm{BJ}}_n \;:=\; n \sup_{t\in(0,1)} K(\hat{G}_n(t),t),$$

where

$$K(s,t) \;:=\; s\log\frac{s}{t} + (1-s)\log\frac{1-s}{1-t}$$

for $s \in [0,1]$ and $t \in (0,1)$. This leads to an alternative confidence band for F: With probability at least $1-\alpha$,

$$nK(\hat{F}_n(x),F(x)) \;\le\; \kappa^{\mathrm{BJ}}_{n,\alpha} \quad \text{for all } x\in\mathbb{R}. \tag{1}$$

As shown by Jager and Wellner (2007), the asymptotic distribution of T^{BJ}_n remains the same if one replaces $K(s,t)$ by a more general function; in particular, one may interchange its two arguments. Moreover,

$$\kappa^{\mathrm{BJ}}_{n,\alpha} \;=\; \log\log(n) + 2^{-1}\log\log\log(n) + O(1).$$

From this one can deduce that (1) leads to confidence intervals with length at most

$$2\big(2\gamma_n F(x)(1-F(x))\big)^{1/2} + 2\gamma_n \quad \text{where} \quad \gamma_n := \frac{\kappa^{\mathrm{BJ}}_{n,\alpha}}{n} = (1+o(1))\frac{\log\log n}{n},$$

uniformly in $x\in\mathbb{R}$. Hence they are substantially shorter than the Kolmogorov-Smirnov intervals for $F(x)$ close to 0 or 1. But in the central region, i.e. when $F(x)$ is bounded away from 0 and 1, they are of width $O(n^{-1/2}(\log\log n)^{1/2})$ rather than $O(n^{-1/2})$. An obvious goal is to refine these methods and combine the benefits of the Kolmogorov-Smirnov and Berk-Jones confidence bands.

A key for our new procedures is a suitable variant of the law of the iterated logarithm (LIL). In what follows we consider the logistic function $\ell:\mathbb{R}\to(0,1)$,

$$\ell(x) \;=\; \frac{e^x}{1+e^x} \;=\; \frac{1}{e^{-x}+1}.$$

Further let $C, D : (0,1)\to[0,\infty)$ be given by

$$\begin{aligned} C(t) \;&:=\; \log\log\frac{e}{4t(1-t)} \;=\; \log\log\Big(\frac{e}{1-(2t-1)^2}\Big) \;\ge\; 0,\\ D(t) \;&:=\; \log(1+C(t)^2) \;\ge\; 0. \end{aligned}$$

Note that $C(t)=C(1-t)$, $D(t)=D(1-t)$, and, as $t\downarrow 0$,

$$\begin{aligned} C(t) \;&=\; \log\log(1/t) + O\big(\log(1/t)^{-1}\big),\\ D(t) \;&=\; 2\log\log\log(1/t) + O\big((\log\log(1/t))^{-1}\big). \end{aligned}$$

Note also that

$$\lim_{t\to1/2}\frac{C(t)}{(2t-1)^2} \;=\; \lim_{t\to1/2}\frac{D(t)}{(2t-1)^4} \;=\; 1.$$

Now let $X=(X(t))_{t\in\mathcal{T}}$ be a nonnegative stochastic process on a set $\mathcal{T}\subset(0,1)$. The following general condition plays a crucial role:

Local uniform sub-exponentiality (LUSE). There exist a real constant $M \geq 1$ and a non-increasing funtion $L : [0, \infty) \to [0, 1]$ such that $L(c) = 1 - O(c)$ as $c \downarrow 0$, and

$$\Pr\Big(\sup_{t \in [\ell(a), \ell(a+c)] \cap \mathcal{T}} X(t) > \eta\Big) \;\leq\; M \exp(-L(c)\eta) \tag{2}$$

for arbitrary $a \in \mathbb{R}$, $c \geq 0$ and $\eta \in \mathbb{R}$.

Theorem 1. Suppose that X satisfies LUSE. For arbitrary $\nu > 1$ and $L_o \in (0, 1)$ there exists a real constant $M_o \geq 1$ depending only on M, $L(\cdot)$, ν and L_o such that

$$\Pr\Big(\sup_{t \in \mathcal{T}}\big(X(t) - C(t) - \nu D(t)\big) > \eta\Big) \;\leq\; M_o \exp(-L_o \eta) \quad \text{for arbitrary } \eta \geq 0.$$

Example 1. Let $X(t) = \mathbb{U}(t)^2/(2t(1-t))$ with standard Brownian bridge $\mathbb{U}$ on $\mathcal{T} = (0, 1)$. Then LUSE is satisfied with $M = 2$ and $L(c) = e^{-c}$.

Example 2. Let $X_n(t) = nK(\hat{G}_n(t), t)$. Then LUSE is satisfied with $M = 2$ and $L(c) = e^{-c}$. This leads to the following new goodness-of-fit test: The null hypothesis that F is equal to a given continuous distribution function F_o is rejected at level α if the test statistic

$$T_{n,\nu}(F_o) \;:=\; \sup_{x \in \mathbb{R}}\big(nK(\hat{F}_n(x), F_o(x)) - C(F_o(x)) - \nu D(F_o(x))\big)$$

exceeds the $(1-\alpha)$-quantile of $\sup_{(0,1)}(X_n - C - \nu D)$. This test has high power for a variety of problems. To verify this, the following inequality for K is crucial:

$$K(x, t) \;\leq\; c \quad \text{implies that} \quad |x - t| \;\leq\; \begin{cases} \sqrt{2c\,x(1-x)} + c, \\ \sqrt{2c\,t(1-t)} + c. \end{cases}$$

Example 3. Let $\mathcal{T}_n = \{t_{n1}, t_{n2}, \ldots, t_{nn}\}$ with $t_{ni} := i/(n+1)$. Further let $U_{n:1} < U_{n:2} < \cdots < U_{n:n}$ be the order statistics of $U_1, U_2, \ldots, U_n$. Now we define $\tilde{X}_n(t_{ni}) = (n+1)K(t_{ni}, U_{n:i})$. Then LUSE is satisfied with $M = 2$ and $L(c) = e^{-c}$. This leads to a confidence band for F: Let $\tilde{\kappa} = \tilde{\kappa}_{n,\nu,\alpha}$ be the $(1-\alpha)$-quantile of $\sup_{\mathcal{T}_n}(\tilde{X}_n - C - \nu D)$. Further let $-\infty = X_{n:0} < X_{n:1} \leq X_{n:2} \leq \cdots \leq X_{n:n} < X_{n:n+1} = \infty$ be the order statistics of $X_1, X_2, \ldots, X_n$. Then with probability at least $1 - \alpha$, the following is true: For $0 \leq j \leq n$ and $X_{n:j} \leq x < X_{n:j+1}$,

$$F(x) \;\in\; [a_{nj}, b_{nj}],$$

where $a_{n0} := 0$, $b_{nn} := 1$ and

$$a_{nj} := \min\big\{u \in [0,1] : nK(t_{nj}, u) \leq C(t_{nj}) + \nu D(t_{nj}) + \tilde{\kappa}\big\} \quad \text{if } j > 0,$$

$$b_{nj} := \max\big\{u \in [0,1] : nK(t_{n,j+1}, u) \leq C(t_{n,j+1}) + \nu D(t_{n,j+1}) + \tilde{\kappa}\big\} \quad \text{if } j < n.$$

It turns out that this confidence band is asymptotically equivalent to Owen's (1995) band in the tail regions but substantially more accurate in the central region. Its maximal width is of order $O(n^{-1/2})$.

References

[1] R. H. Berk, D. H. Jones, *Goodness-of-fit test statistics that dominate the Kolmogorov statistics*, Z. Wahrsch. Verw. Gebiete, 47 (1979), 47–59.
[2] L. Dümbgen, Jon A. Wellner, *Confidence bands for a distribution function: A new look at the law of the iterated logarithm*, Preprint (2014) (arxiv:1402.2918).
[3] A. B. Owen, *Nonparametric likelihood confidence bands for a distribution function*, J. Amer. Statist. Assoc., 90 (1995), 516–521.

An open problem on strongly consistent learning of the best prediction for Gaussian processes

László Györfi
(joint work with Alessio Sancetta)

Let $\{Y_n\}_{-\infty}^{\infty}$ be a stationary, ergodic, mean zero Gaussian process. The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$. It is an open problem whether it is possible to learn the best predictor from the past data in a strongly consistent way, i.e., whether there exists a prediction rule g such that

$$\lim_{n\to\infty} (\mathbf{E}\{Y_n \mid Y_1, \dots, Y_{n-1}\} - g_n(Y_1, \dots, Y_{n-1})) = 0 \quad \text{almost surely} \tag{1}$$

for all stationary and ergodic Gaussian processes.

In [1] we summarized some positive and negative findings in this respect.

References

[1] L. Györfi, A. Sanchetta, *An open problem on strongly consistent learning of the best prediction for Gaussian processes*, in *Topics in Nonparametric Statistics. Proceedings of the First Conference of the International Society of Nonparametric Statistics*, ed. by M. Akritas, S. N. Lahiri and D. Politis, Springer, Heidelberg, (2014). http://www.cs.bme.hu/∼gyorfi/gysafinal.pdf

Inverse Problems with Poisson Data: Pioneering contributions of L. Cavalier and recent developments

Thorsten Hohage

We consider inverse problems described by an operator equation

$$F(u) = g \tag{1}$$

with a possibly nonlinear injective forward operator $F : D(F) \subset \mathcal{X} \to L^1(\mathbb{M})$ where $\mathbb{M} \subset \mathbb{R}^d$ is a smooth manifold with data. Let $u^\dagger \in D(F)$ denote the exact solution, $g^\dagger := F(u^\dagger)$, and suppose that $F(u) \geq 0$ for all $u \in D(F)$. We assume that data $\tilde{G}_t = \sum_{i=1}^{N} \delta_{x_i}$ are drawn from a Poisson process with density $tF(u^\dagger)$, where $t > 0$ can typically be interpreted as an exposure time, and study the convergence of estimators as $t \to \infty$. It will be convenient to define $G_t := \tilde{G}_t/t$.

Such inverse problems typically occur in photonic imaging applications such as Positron Emission Tomography (PET), confocal fluorescence microscopy, coherent x-ray imaging, and inverse scattering problems with low energy densities.

To our knowledge the setup above was first considered in the paper [2] by Cavalier & Koo. In this work the authors studied PET, where the forward operator F is given by the Radon transform. They obtained the following results: On projected Besov balls $\mathcal{F}^s_{pq} := \{u \in B^s_{pq} : u \geq 0, \|u\|_{spq} \leq R\}$ with $p, q \in [1, \infty]$ and $s > 2/p$ the minimax rate for this problem is of order $O(t^{-s/(2s+3)})$. For known smoothness this rate can be obtained by thresholding of empirical vagulette coefficients in a wavelet-vagulette decomposition of the Radon transform F. For unknown smoothness the rate $O((t/\ln t)^{-s/(2s+3)})$ can be obtained. Later the results in [2] were extended to more general linear forward operators by Antoniadis & Bigot [1].

In the following we will study regularization of nonlinear inverse problems (1) by nonlinear Tikhonov (or penalized maximum likelihood) regularization

$$\widehat{u}_\alpha \in \operatorname*{argmin}_{u \in \mathfrak{B}} \left[\mathcal{S}(G_t, F(u)) + \alpha \|u - u_0\|^2_{\mathcal{X}}\right]. \tag{2}$$

Here we assume for simplicity that $\mathcal{X}$ is a Hilbert space. $u_0 \in \mathcal{X}$ is some initial guess (e.g. $u_0 = 0$). $\|u - u_0\|^2_{\mathcal{X}}$ can be replaced by more general convex penalty functionals $\mathcal{R}(u)$. Then one obtains convergence with respect to the Bregman distance of $\mathcal{R}$.

A first idea for the choice of $\mathcal{S}$ is the log-likelihood functional $\mathcal{S}_0(G_t, g) = \int_{\mathbb{M}} g\, dx - \int_{\mathbb{M}} \ln(g)\, \mathrm{d}G_t$. Note that $\mathbb{E}[\mathcal{S}_0(G_t, g) - \mathcal{S}_0(G_t, g^\dagger)] = \mathrm{KL}(g^\dagger, g)$ (the Kullback-Leibler divergence) and

$$\mathcal{S}_0(G_t, g) - \mathcal{S}_0(G_t, g^\dagger) - \mathrm{KL}(g^\dagger, g) = \int -\ln \frac{g}{g^\dagger}(dG_t - g^\dagger dx)$$

Our analysis relies on uniform estimates of the right hand side by concentration inequalities, which are not applicable unless $\|g/g^\dagger\|_\infty$ is uniformly bounded in g. Therefore, we introduce a shift parameter $\tau \geq 0$, define $\mathcal{T}(g^\dagger, g) := \mathrm{KL}(g^\dagger + \tau, g + \tau)$ and $\mathcal{S}(G_t, g) = \mathcal{S}_\tau(G_t, g) = \int_{\mathbb{M}} g\, dx - \int_{\mathbb{M}} \ln(g + \tau)\, (\mathrm{d}G_t + \tau dx)$ such that

$$\mathcal{S}_\tau(G_t, g) - \mathcal{S}_\tau(G_t, g^\dagger) - \mathcal{T}(g^\dagger, g) = \int -\ln \frac{g + \tau}{g^\dagger + \tau}(dG_t - g^\dagger dx).$$

Now if

$$\sup_{u \in \mathfrak{B}} \|F(u)\|_{H^s} < \infty \tag{3}$$

and $\tau > 0$, then it can be shown based on results in [7] that there exists $C > 0$ such that

$$\mathbb{P}\left[\sup_{g \in F(\mathfrak{B})} \left| \int \ln \frac{g + \tau}{g^\dagger + \tau}(dG_t + \tau dx) \right| \geq \frac{\rho}{\sqrt{t}}\right] \leq \exp\left(-\frac{\rho}{C}\right) \tag{4}$$

for all $t, \rho \geq 1$ (see [8]).

To show rates of convergence for ill-posed problem some sort of smoothness condition has to be imposed. Since first suggested (with $\varphi(x) = x$) in [5], such conditions are often formulated in the form of variational inequalities: We assume

that there exists $\beta \in (0,1]$ and a concave, increasing function $\varphi : [0,\infty) \to \mathbb{R}$ with $\varphi(0) = 0$ such that for all $u \in \mathfrak{B}$

$$\beta\|u - u^\dagger\|^2 \le \|u - u_0\|^2 - \|u^\dagger - u_0\|^2 + \varphi\left(\mathcal{T}(F(u^\dagger), F(u))\right). \tag{5}$$

We point out that a Hölder source condition $u^\dagger = (F^*F)^\nu w$ with $\nu \in (0, 1/2]$ for a bounded linear operator F in Hilbert spaces implies a variational source condition with $\varphi(t) = ct^{\frac{2\nu}{2\nu+1}}$. Moreover, as opposed to standard spectral source condition $u^\dagger = \psi(F^*F)w$, variational source conditions in Hilbert spaces (or equivalent concepts) even allow sharp converse results [3].

Theorem: (see [8]) *Suppose that* (3) *and* (5) *hold true. Then with the a-priori parameter choice rule* $\frac{-1}{\alpha} \in \partial(-\varphi)(t^{-1/2})$ *the risk is bounded by*

$$\mathbb{E}\left[\|\widehat{u}_\alpha - u^\dagger\|^2\right] = \mathcal{O}\left(\varphi\left(t^{-1/2}\right)\right) \qquad t \to \infty. \tag{6}$$

Without a-priori knowledge of the function φ *the Lepskiĭ-type balancing principle*

$$\alpha_{\mathrm{bal}} := \max\left\{j \in \mathbb{N} : \|u_{\alpha_j} - u_{\alpha_k}\| \le 8r^{-j/2} \text{ for } k = 0, \dots, j-1\right\}$$

with $\alpha_j = \alpha_0 r^{-j}$, $r > 1$ *leads to the risk bound*

$$\mathbb{E}\left[\|\widehat{u}_{\alpha_{\mathrm{bal}}} - u^\dagger\|^2\right] = \mathcal{O}\left(\varphi\left(\ln(t)t^{-1/2}\right)\right) \qquad t \to \infty. \tag{7}$$

A disadvantage of Tikhonov regularization is the fact that the objective funtional is non-convex in general, and it may have many local minima. An alternative is to linearize the operator and use a Newton-type method: Choose $\alpha_k = \alpha_0\rho^k$ for some $\rho \in (0,1)$ and set

$$u_{k+1} \in \operatorname*{argmin}_{u \in \mathfrak{B}} \left[\mathcal{S}\left(G_t, F'[u_k](u - u_k) + F(u_k)\right) + \alpha_k \mathcal{R}(u)\right].$$

If $\mathcal{S}$ and $\mathcal{R}$ are convex, a convex optimization problem has to be solved in each Newton step. Here the choice of the stopping index corresponds to the choice of α. Under an additional assumption on the local approximation quality of F' (a tangential cone condition) we can show similar results as for Tikhonov regularization (see [6]).

References

[1] A. Antoniadis and J. Bigot. Poisson inverse problems. *Ann. Statist.*, 34(5):2132–2158, 2006.

[2] L. Cavalier and J.-Y. Koo. Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. *IEEE Trans. Inform. Theory*, 48(10):2794–2802, 2002.

[3] J. Flemming, B. Hofmann and P. Mathé. Sharp converse results for the regularization error using distance functions. Inverse Problems, 27(18):025006, 2011

[4] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Probl.*, 26(11):115014, 2010.

[5] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl.*, 23(3):987–1010, 2007.

[6] T. Hohage and F. Werner. Iteratively regularized Newton-type methods for general data misfit functionals and applications to Poisson data. *Numer. Math.*, 123(4):745–779, 2013.

[7] P. Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Rel.*, 126(1):103–153, 2003.
[8] F. Werner and T. Hohage. Convergence rates in expectation for Tikhonov-type regularization of Inverse Problems with Poisson data. *Inverse Probl.*, 28(10):104004, 2012.

Sparse high-dimensional varying coefficient model: non-asymptotic minimax study

Olga Klopp
(joint work with Marianna Pensky)

One of the fundamental tasks in statistics is to characterize the relationship between a set of covariates and a response variable. In the present work we study the varying coefficient model which is commonly used for describing time-varying covariate effects. It provides a more flexible approach than the classical linear regression model and is often used to analyze the data measured repeatedly over time.

Let $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \dots, n$ be sampled independently from the varying coefficient model

$$Y = \mathbf{W}^T \mathbf{f}(t) + \sigma\xi. \tag{1}$$

Here the noise variables ξ_i are independent and σ is known, $\mathbf{W} \in \mathbb{R}^p$ are random vectors of predictors, $\mathbf{f}(\cdot) = (f_1(\cdot), \dots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients and $t \in [0, 1]$ is a random variable with the unknown density function g. We suppose that $\mathbf{W}$ and t are independent. The goal is to estimate vector function $f(\cdot)$ on the basis of observations $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \dots, n$.

Since its introduction by Cleveland, Grosse and Shyu [1] and Hastie and Tibshirani [3] many methods for estimation and inference in the varying coefficient model have been developed. Existing methods typically provide asymptotic evaluation of the precision of the estimation procedure under the assumption that the number of observations tends to infinity and is larger than the dimension of the problem. Recently few authors consider still asymptotic but high-dimensional approach to the problem. Wei *et al.* [7] applied group Lasso for variable selection, while Lian [5] used extended Bayesian information criterion. Fan *et al.* [2] applied nonparametric independence screening. Their results were extended by Lian and Ma [6] to include rank selection in addition to variable selection.

One important aspect that has not been well studied in the existing literature is the non-asymptotic approach to the estimation, prediction and variables selection in the varying coefficient model. Some interesting questions arise in this non-asymptotic setting. One of them is the fundamental question of the minimax optimal rates of convergence. The minimax risk characterizes the essential statistical difficulty of the problem. It also captures the interplay between different parameters in the model. To the best of our knowledge, our work presents the first *non-asymptotic minimax study* of the sparse heterogeneous varying coefficient model.

Modern technologies produce very high dimensional data sets and, hence, stimulate an enormous interest in variable selection and estimation under a sparse scenario. In the present work, we consider the case when the solution is sparse, in particular, only few of the covariates are present and only some of them are time dependent. We consider a quite flexible and realistic scenario where the time dependent covariates possibly have different degrees of smoothness and may be spatially inhomogeneous.

In order to estimate $\mathbf{f}$, following Klopp and Pensky [4], we expand it over a basis $(\phi_l(\cdot)), l = 0, 1, \ldots, \infty$, in $L_2([0,1])$ with $\phi_0(t) = 1$. Expansion of the functions $f_j(\cdot)$ over the basis, for any $t \in [0,1]$, yields

$$f_j(t) = \sum_{l=0}^{L} a_{jl}\phi_l(t) + \rho_j(t) \quad \text{with} \quad \rho_j(t) = \sum_{l=L+1}^{\infty} a_{jl}\phi_l(t). \tag{2}$$

If $\phi(\cdot) = (\phi_0(\cdot), \ldots, \phi_L(\cdot))^T$ and $\mathbf{A}_0$ denotes a matrix of coefficients with elements $A_0^{(l,j)} = a_{jl}$, then relation (2) can be re-written as $\mathbf{f}(t) = \mathbf{A}_0^T\phi(t) + \rho(t)$, where $\rho(t) = (\rho_1(t), \cdots, \rho_p(t))^T$. Combining formulae (1) and (2), we obtain the following model for observations $(\mathbf{W}_i, t_i, Y_i)$, $i = 1, \ldots, n$:

$$Y_i = \mathrm{Tr}(\mathbf{A}_0^T\phi(t_i)\mathbf{W}_i^T) + \mathbf{W}_i^T\rho(t_i) + \sigma\xi_i, \quad i = 1, \ldots, n. \tag{3}$$

Below, we reduce the problem of estimating vector function $\mathbf{f}$ to estimating matrix $\mathbf{A}_0$ of coefficients of $\mathbf{f}$.

We construct a minimax optimal estimator using the block Lasso which can be viewed as a version of the group LASSO. However, unlike in group LASSO, where the groups occur naturally, the blocks in block LASSO are driven by the need to reduce the variance as it is done, for example, in block thresholding. In particular, for each function f_j, $j = 1, \cdots, p$, we divide its coefficients into $M + 1$ different groups where group zero contains only coefficient a_{j0} for the constant function $\phi_0(t) = 1$ and M groups of size $d \approx \log n$ where $M = L/d$. We denote $\mathbf{a}_{j0} = a_{j0}$ and $\mathbf{a}_{jl} = (a_{j,d(l-1)+1}, \cdots, a_{j,dl})^T$ the sub-vector of coefficients of function f_j in block l, $l = 1, \cdots, M$. Let K_l be the subset of indices associated with $\mathbf{a}_{jl}$. We impose block norm on matrix $\mathbf{A}$ as follows

$$\|\mathbf{A}\|_{\mathrm{block}} = \sum_{j=1}^{p}\sum_{l=0}^{M} \|\mathbf{a}_{jl}\|_2. \tag{4}$$

Observe that $\|\mathbf{A}\|_{\mathrm{block}}$ indeed satisfies the definition of a norm and is a sum of absolute values of coefficients a_{j0} of functions f_j and l_2 norms for each of the block vectors of coefficients $\mathbf{a}_{jl}$, $j = 1, \cdots, p$, $l = 1, \cdots, M$.

We construct an estimator $\hat{\mathbf{A}}$ of $\mathbf{A}_0$ as a solution of the following convex optimization problem

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ n^{-1}\sum_{i=1}^{n} \left(Y_i - \mathrm{Tr}(\mathbf{A}^T\phi(t_i)\mathbf{W}_i^T)\right)^2 + \delta\|\mathbf{A}\|_{\mathrm{block}} \right\}, \tag{5}$$

where the value of δ is the regularization parameter.

Our estimator does not require the knowledge which of the covariates are indeed present and which are time dependent. It adapts to sparsity, to heterogeneity of the time dependent covariates and to their possibly spatial inhomogeneous nature. In order to ensure the optimality, we derive minimax lower bounds for the risk and show that our estimator attains those bounds within a constant (if all time-dependent covariates are spatially homogeneous) or logarithmic factor of the number of observations. The analysis is carried out under the flexible assumption that the noise variables are sub-Gaussian. In addition, it does not require that the elements of the dictionary are uniformly bounded.

References

[1] Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991) Local regression models. *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309-376. Wadsworth and Books, Pacific Grove.

[2] Fan, J., Ma, Y., and Dai, W. (2013) Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models. `arxiv:1303.0458v1`

[3] Hastie, T.J. and Tibshirani, R.J. (1993) Varying-coefficient models. *J. Roy. Statist. Soc. B.* (Chambers, J.M. and Hastie, T.J., eds), **55** 757-796.

[4] Klopp, O., Pensky, M. (2013) Non-asymptotic approach to varying coefficient model. *Electronic Journal of Statistics*, **7**, 454-479.

[5] Lian, H. (2012) Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, **22**, 1563-1588.

[6] Lian, H., and Ma, S. (2013) Reduced-rank Regression in Sparse Multivariate Varying-Coefficient Models with High-dimensional Covariates. `arxiv:1309.6058v1`

[7] Wei, F., Huang, J. and Li, G. (2011) Variable Selection and Estimation in High-Dimensional Varying-Coefficient Models. *Statistica Sinica.*, **21**, 1515-1540.

The restricted eigenvalue assumption under weak moment assumption

Guillaume Lecué

(joint work with Shahar Mendelson)

We prove that iid random vectors that satisfy a rather weak moment assumption can be used as measurement vectors in Compressed Sensing. In many cases, the moment assumption suffices to ensure that the number of measurements required for exact reconstruction is the same as the best possible estimate – exhibited by a random gaussian matrix. In Compressed Sensing (see, e.g., [5] and [8]), one observes linear measurements $y_i = \langle X_i, x_0 \rangle$, $i = 1, ..., N$ of an unknown vector $x_0 \in \mathbb{R}^n$, and the goal is to identify x_0 using those measurements.

Given the measurements matrix $\Gamma = N^{-1/2} \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$, a possible recovery procedure is the basis pursuit algorithm, defined by

$$\hat{x} \in \operatorname{argmin}\big(\|t\|_1 : \Gamma t = \Gamma x_0 \big).$$

A well known question is to identify conditions on the vectors $X_1, ..., X_N$ that ensure that if x_0 is s-sparse, that is, if it is supported on at most s coordinates, the unique minimizer of the basis pursuit algorithm is x_0 itself. The matrix Γ satisfies

the exact reconstruction property in Σ_s, the set of all s-sparse vectors in $\mathbb{R}^n$, if every $x_0 \in \Sigma_s$ has this property.

A standard choice of a measurements matrix Γ is when $X_1, ..., X_N$ are independent, isotropic and L-subgaussian random vectors. Recall that a random vector X in $\mathbb{R}^n$ is isotropic if for every $t \in \mathbb{R}^n$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$, and it is L-subgaussian if for every $t \in \mathbb{R}^n$ and every $p \geq 2$, $\|\langle X, t\rangle\|_{L_p} \leq L\sqrt{p}\|\langle X, t\rangle\|_{L_2}$.

One may show that if the X_i's are random vectors that are independent, isotropic and L-subgaussian, then with high probability Γ satisfies the exact reconstruction property for s-sparse vectors as long as $N \gtrsim s\log(en/s)$ [10], and this number of measurements cannot be improved (see Proposition 2.2.18 in [7]).

The reason behind this result, and many others like it, is that isotropic subgaussian matrices act on Σ_s in an isomorphic way, when $N \gtrsim s\log(en/s)$.

Such a property is called the *Restricted isometry property* (RIP) (see, for example [4, 6, 11]). A matrix Γ satisfies the RIP in Σ_s if for every $t \in \Sigma_s$,

$$(1-\delta)\|t\|_2 \leq \|\Gamma t\|_2 \leq (1+\delta)\|t\|_2,$$

for some fixed $0 < \delta < 1$.

Proving the RIP for subgaussian matrices uses the fact that tails of linear functionals $\langle X, t\rangle$ decay faster than the corresponding gaussian variable. Thus, it seemed natural to ask whether the same type of estimates hold in cases where linear functionals exhibit a slower decay – for example, when X is sub-exponential, and the linear functionals satisfy that $\|\langle X, t\rangle\|_{L_p} \leq Lp\|\langle X, t\rangle\|_{L_2}$ for every $t \in \mathbb{R}^n$ and every $p \geq 2$.

Proving the RIP for a sub-exponential ensemble is a much harder task than for subgaussian ensembles (cf. [3]). Moreover, the RIP does not exhibit the same behaviour as in the gaussian case. Indeed, one may show that for sub-exponential ensembles, RIP holds with high probability only when $N \gtrsim s\log^2(en/s)$, and this estimate is optimal as can be seen when X has independent, symmetric exponential random variables as coordinates [3].

On the other hand, the result in [9] (see Chapter 7 there) shows that exact reconstruction can still be achieved by isotropic sub-exponential measurement vectors when $N \gtrsim s\log(en/s)$ – the same number of measurements needed for the gaussian ensemble.

Clearly, this estimate cannot be based on the RIP, and one may ask whether weaker tail assumptions on the measurement vectors may still lead to exact recovery with the 'gaussian' number of measurements.

The main result presented here is precisely in this direction:

Theorem A. There exist absolute constants c_0, c_1 and c_2 and for every $\alpha \geq 1/2$ there exists a constant $c_3(\alpha)$ that depends only on α for which the following holds. Let $X = (x_i)_{i=1}^n$ be a random vector on $\mathbb{R}^n$ such that

(1) There are $\kappa_1, \kappa_2, w > 1$ that satisfy that for every $1 \leq j \leq n$, $\|x_j\|_{L_2} = 1$, and for $p = \kappa_2\log(wn)$, $\|x_j\|_{L_p} \leq \kappa_1 p^\alpha$.

(2) There are $u, \beta > 0$ that satisfy for every $t \in \Sigma_s \cap S^{n-1}$,

$$P\left(|\langle X, t\rangle| > u\right) \geq \beta.$$

If

$$N \geq c_0 \max\left\{ s\log(en/s), (c_3(\alpha)\kappa_1)^2(\kappa_2 \log(wn))^{\max\{2\alpha-1,1\}} \right\}$$

and $X_1, ..., X_N$ are independent copies of X, then, with probability at least $1 - 2\exp(-c_1\beta^2 N) - 1/w^{\kappa_2} n^{\kappa_2 - 1}$, $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$ satisfies the exact reconstruction property in Σ_{s_1} for $s_1 = c_2 u^2 \beta s$.

It follows from Theorem A, that a random matrix with iid centered entries that have variance 1 and an L_p moment bounded by p for $p = 2\log n$ can be used as a measurement matrix, and just as in the gaussian case, requires only $N \gtrsim s\log(en/s)$ measurements.

Just as noted for sub-exponential ensembles, Theorem A cannot be proved using the RIP, and its proof must take a different path.

Note that we proved a stronger result: under the same assumptions as in Theorem A, the measurement matrix Γ satisfies the Restricted eigenvalue assumption as introduced in [2].

References

[1] G. Lecué and S. Mendelson *Compressed sensing under weak moment assumption*, Arxiv.

[2] P.J. Bickel, Y. Ritov and A.B. Tsybakov *Simultaneous analysis of Lasso and Dantzig selector.*, Annals of Statistics, v.37, n.4, 1705-1732 (2009).

[3] Radosław Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constr. Approx.*, 34(1):61–88, 2011.

[4] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[5] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[6] Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[7] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2012.

[8] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[9] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[10] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[11] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2008.

Adaptive estimation over anisotropic functional classes via oracle approach

Oleg Lepski

We address the problem of adaptive minimax estimation in white gaussian noise model under $\mathbb{L}_p$–loss, $1 \le p \le \infty$, on the anisotropic Nikolskii classes. We present the estimation procedure based on a new data-driven selection scheme from the family of kernel estimators with varying bandwidths. For proposed estimator we establish so-called $\mathbb{L}_p$-norm oracle inequality and use it for deriving minimax adaptive results. We prove the existence of rate-adaptive estimators and fully characterize behavior of the minimax risk for different relationships between regularity parameters and norm indexes in definitions of the functional class and of the risk. In particular some new asymptotics of the minimax risk are discovered including necessary and sufficient conditions for existence a uniformly consistent estimator. Consider the sequence of statistical experiments (called gaussian white noise model) generated by the observation $X^\varepsilon = \big\{Y_\epsilon(g),\ g \in \mathbb{L}_2\big(\mathbb{R}^d, \nu_d\big)\big\}_\varepsilon$ where

$$Y_\varepsilon(g) = \int f(t)g(t)\nu_d(\mathrm{d}t) + \varepsilon \int g(t)W(\mathrm{d}t). \tag{1}$$

Here $\varepsilon \in (0,1)$ is understood as the noise level which is usually supposed sufficiently small.

The goal is to recover unknown signal f from observation X^ε on a given cube $(-b,b)^d,\ b > 0$. The quality of an estimation procedure will be described by $\mathbb{L}_p$-risk, $1 \le p \le \infty$, defined in (2) below and as an estimator we understand any X^ε-measurable Borel function belonging to $\mathbb{L}_p\big(\mathbb{R}^d, \nu_d\big)$. Without loss of generality and for ease of the notation we will assume that functions to be estimated vanish outside $(-b,b)^d$.

Thus, for any estimator $\tilde{f}_\varepsilon$ and any $f \in \mathbb{L}_p\big(\mathbb{R}^d, \nu_d\big) \cap \mathbb{L}_2\big(\mathbb{R}^d, \nu_d\big)$ we define its $\mathbb{L}_p$-risk as

$$\mathcal{R}_\varepsilon^{(p)}\big[\tilde{f}_\varepsilon; f\big] = \Big\{\mathbb{E}_f^{(\varepsilon)}\Big(\big\|\tilde{f}_\varepsilon - f\big\|_p^q\Big)\Big\}^{\frac{1}{q}},\ \ q \ge 1. \tag{2}$$

Here and later $\|\cdot\|_p, 1 \le p \le \infty$, stands for $\|\cdot\|_{p,(-b,b)^d}$ and $\mathbb{E}_f^{(\varepsilon)}$ denote the mathematical expectation with respect to the probability law of X^ε.

Let $\mathbb{F}$ be a given subset of $\mathbb{L}_p\big(\mathbb{R}^d, \nu_d\big) \cap \mathbb{L}_2\big(\mathbb{R}^d, \nu_d\big)$. For any estimator $\tilde{f}_\varepsilon$ define its *maximal risk* by $\mathcal{R}_\varepsilon^{(p)}\big[\tilde{f}_\varepsilon; \mathbb{F}\big] = \sup_{f\in\mathbb{F}} \mathcal{R}_\varepsilon^{(p)}\big[\tilde{f}_\varepsilon; f\big]$ and its *minimax risk* on $\mathbb{F}$ is given by

$$\phi_\varepsilon(\mathbb{F}) := \inf_{\tilde{f}_\varepsilon} \mathcal{R}_\varepsilon^{(p)}\big[\tilde{f}_\varepsilon; \mathbb{F}\big]. \tag{3}$$

Here infimum is taken over all possible estimators. An estimator whose risk is proportional to $\phi_\varepsilon(\mathbb{F})$ is called minimax on $\mathbb{F}$.

Let $\big\{\mathbb{F}_\vartheta, \vartheta \in \Theta\big\}$ be the collection of subsets of $\mathbb{L}_p\big(\mathbb{R}^d, \nu_d\big) \cap \mathbb{L}_2\big(\mathbb{R}^d, \nu_d\big)$, where ϑ is a nuisance parameter which may have very complicated structure.

The problem of adaptive estimation can be formulated as follows: *is it possible to construct a single estimator $\hat{f}_\varepsilon$ which would be simultaneously minimax on each class $\{\mathbb{F}_\vartheta,\ \vartheta\in\Theta\}$, i.e.*

$$\mathcal{R}_\varepsilon^{(p)}\big[\hat{f}_\varepsilon;\mathbb{F}_\vartheta\big]\sim\phi_\varepsilon(\mathbb{F}_\vartheta),\ \varepsilon\to 0,\ \ \forall\vartheta\in\Theta?$$

We refer to this question as *the problem of adaptive estimation over the scale of* $\{\mathbb{F}_\vartheta,\ \vartheta\in\Theta\}$. If such estimator exists we will call it optimally or rate-adaptive.

In the present paper we will be interested in adaptive estimation over the scale

$$\mathbb{F}_\vartheta=\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big),\ \vartheta=\big(\vec{\beta},\vec{r},\vec{L}\big),$$

where $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ is an anisotropic Nikolskii class.

Let $(\mathbf{e}_1,\dots,\mathbf{e}_d)$ denote the canonical basis of $\mathbb{R}^d$. For function $g:\mathbb{R}^d\to\mathbb{R}^1$ and real number $u\in\mathbb{R}$ define *the first order difference operator with step size u in direction of the variable x_j* by

$$\Delta_{u,j}g(x)=g(x+u\mathbf{e}_j)-g(x),\quad j=1,\dots,d.$$

By induction, the k-th order difference operator with step size u in direction of the variable x_j is defined as

$$\Delta_{u,j}^k g(x)=\Delta_{u,j}\Delta_{u,j}^{k-1}g(x)=\sum_{l=1}^{k}(-1)^{l+k}\binom{k}{l}\Delta_{ul,j}g(x). \tag{4}$$

Definition 1. *For given vectors $\vec{r}=(r_1,\dots,r_d)$, $r_j\in[1,\infty]$, $\vec{\beta}=(\beta_1,\dots,\beta_d)$, $\beta_j>0$, and $\vec{L}=(L_1,\dots,L_d)$, $L_j>0$, $j=1,\dots,d$, we say that function $g:\mathbb{R}^d\to\mathbb{R}^1$ belongs to the anisotropic Nikol'skii class $\bar{\mathbb{N}}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ if*

(i) $\|g\|_{r_j,\mathbb{R}^d}\le L_j$ *for all* $j=1,\dots,d$;

(ii) *for every* $j=1,\dots,d$ *there exists natural number* $k_j>\beta_j$ *such that*

$$\Big\|\Delta_{u,j}^{k_j}g\Big\|_{r_j,\mathbb{R}^d}\le L_j|u|^{\beta_j},\quad\forall u\in\mathbb{R},\quad\forall j=1,\dots,d. \tag{5}$$

Recall that the consideration of white gaussian noise model requires $f\in\mathbb{L}_2\big(\mathbb{R}^d\big)$ that is not always guaranteed by $f\in\bar{\mathbb{N}}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$. So, later on we will study the functional classes $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)=\bar{\mathbb{N}}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)\cap\mathbb{L}_2\big(\mathbb{R}^d\big)$ which we will also call anisotropic Nikol'skii classes.

Let $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ be the anisotropic Nikol'skii functional class. Put

$$\frac{1}{\beta}:=\sum_{j=1}^{d}\frac{1}{\beta_j},\qquad\frac{1}{\omega}:=\sum_{j=1}^{d}\frac{1}{\beta_j r_j},\qquad L_\beta:=\prod_{j=1}^{d}L_j^{1/\beta_j},$$

and define for any $1\le s\le\infty$

$$\tau(s)=1-1/\omega+1/(s\beta),\qquad\kappa(s)=\omega(2+1/\beta)-s.$$

Set finally $p^* = \big[\max_{j=1,\dots,d} r_l\big] \vee p$ and introduce

$$\mathfrak{a} = \begin{cases} \frac{\beta}{2\beta+1}, & \kappa(p) > 0; \\ \frac{1-1/\omega+1/(\beta p)}{2-2/\omega+1/\beta}, & \kappa(p) \le 0,\ \tau(p^*) > 0; \\ \frac{\omega(p^*-p)}{p(p^*-\omega(2+1/\beta)}, & \kappa(p) \le 0,\ \tau(p^*) \le 0,\ p^* > p; \\ 0, & \kappa(p) \le 0,\ \tau(p^*) \le 0;\ p^* = p. \end{cases}$$

$$\delta_\varepsilon = \begin{cases} L_\beta \varepsilon^2, & \kappa(p) > 0; \\ L_\beta \varepsilon^2 |\ln(\varepsilon)|, & \kappa(p) \le 0,\ \tau(p^*) \le 0; \\ L_\beta^{\frac{1-2/p}{\tau(p)}} \varepsilon^2 |\ln(\varepsilon)|, & \kappa(p) \le 0,\ \tau(p^*) > 0. \end{cases}$$

Lower bound of minimax risk

Theorem 1. *Let $q \ge 1$, $L_0 > 0$ and $1 \le p \le \infty$ be fixed. Then for any $\vec{\beta} \in (0,\infty)^d$, $\vec{r} \in [1,\infty]^d$ and $\vec{L} \in [L_0,\infty)^d$ there exists $c > 0$ independent of $\vec{L}$ such that*

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{f}_\varepsilon} \sup_{f \in \mathbb{N}_{\vec{r},d}\left(\vec{\beta},\vec{L}\right)} \delta_\varepsilon^{-\mathfrak{a}} \mathcal{R}_\varepsilon^{(p)}\big[\tilde{f}_\varepsilon; f\big] \ge c,$$

where infimum is taken over all possible estimators.

Let us make several remarks.

1^0. *Case $p^* = p$.* We note that there is no a uniformly consistent estimator over $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ if

$$\tau(p)1_{[2,\infty)}(p) + \kappa(p)1_{[1,2)}(p) \le 0, \tag{6}$$

and this result seems to be new. As it will follow from the next theorem the latter condition is necessary and sufficient for nonexistence of uniformly consistent estimators over $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ under $\mathbb{L}_p$-loss, $1 \le p \le \infty$. In the case of $\mathbb{L}_\infty$-loss, (6) is reduced to $\omega \le 1$ and the similar result was recently proved in [1] for the density model.

2^0. *Case $\kappa(p) \le 0,\ \tau(p^*) \le 0,\ p^* > p$.* The lower bound for minimax risk given in this case by

$$\big(L_\beta \varepsilon^2 |\ln(\varepsilon)|\big)^{\frac{\omega(p^*-p)}{p(p^*-\omega(2+1/\beta)}}$$

is new. It is interesting that the latter case does not appear in the dimension 1 or, more generally, when isotropic Nikolskii classes are considered. Indeed, if $r_l = r$ for all $l = 1,\dots d$, then $p^* > p$ means $r > p$ that, in its turn, implies $\tau(p^*) = \tau(r) = 1 > 0$. It is worth mentioning that we improve in order the lower bound recently found in [1], which corresponds formally to our case $p^* = \infty$.

3^0. *Case $\kappa(p) \le 0,\ \tau(p^*) > 0$.* For the first time the same result was proved in [3] but under more restrictive assumption $\kappa(p) \le 0,\ \tau(\infty) > 0$. Moreover, the dependence of the asymptotics of the minimax risk on $\vec{L}$ was not optimal.

Adaptive upper bound

For any $\ell \in \mathbb{N}^*$ and $L_0 > 0$ set $\Theta = (0,\ell]^d \times [1,\infty]^d \times [L_0,\infty)^d$ and later on we will use the notation $\vartheta \in \Theta$ for the triplet $\big(\vec{\beta},\vec{r},\vec{L}\big)$. Denote $\mathcal{P} = \Theta \times [1,\infty]$ and introduce

$$\mathcal{P}^{\rm consist} = \big\{(\vartheta,p) \in \mathcal{P} : \tau(p)1_{[2,\infty)}(p) + \kappa(p)1_{[1,2)}(p) > 0\big\} \cup \big\{(\vartheta,p) \in \mathcal{P} : p^* > p\big\}.$$

The latter set consists of the class parameters and norm indexes for which a uniform consistent estimation is possible. Introduce $L^* = \min_{j:r_j=p^*} L_j$ and

$$\boldsymbol{\delta}_\varepsilon \quad = \quad \begin{cases} L_\beta \varepsilon^2, & \kappa(p) \ge 0; \\ L_\beta (L^*)^{\frac{1}{\mathfrak{a}}} \varepsilon^2 |\ln(\varepsilon)|, & \kappa(p) \le 0,\ \tau(p^*) \le 0; \\ V_p\big(\vec{L}\big)\varepsilon^2 |\ln(\varepsilon)|, & \kappa(p) \le 0,\ \tau(p^*) > 0. \end{cases}$$

Let $q \ge 1$, $L_0 > 0$ and $\ell > 0$ be fixed. One can construct estimators $\widehat{f}$ and $\hat{f}^{(\rm const)}$ for which the following results hold.

Theorem 2. *1) For any $(\vartheta,p) \in \mathcal{P}^{consist}$ such that $p \in (1,\infty)$, $\vec{r} \in (1,\infty]^d$ and $\kappa(p) \ne 0$ there exists $C > 0$ independent of $\vec{L}$ for which*

$$\limsup_{\varepsilon\to 0} \sup_{f\in\mathbb{N}_{\vec{r},d}\left(\vec{\beta},\vec{L}\right)} \boldsymbol{\delta}_\varepsilon^{-\mathfrak{a}} \mathcal{R}_\varepsilon^{(p)}\big[\widehat{f};f\big] \le C.$$

2) For any $(\vartheta,p) \in \mathcal{P}^{consist}$, $p \in \{1,\infty\}$ there exists $C > 0$ independent of $\vec{L}$ for which

$$\limsup_{\varepsilon\to 0} \sup_{f\in\mathbb{N}_{\vec{r},d}\left(\vec{\beta},\vec{L}\right)} \boldsymbol{\delta}_\varepsilon^{-\mathfrak{a}} \mathcal{R}_\varepsilon^{(p)}\big[\hat{f}^{(const)};f\big] \le C.$$

3) For any $(\vartheta,p) \in \mathcal{P}^{consist}$ such that $p \in (1,\infty)$, $\vec{r} \in (1,\infty]^d$ and $\kappa(p) = 0$ there exists $C > 0$ independent of $\vec{L}$ for which

$$\limsup_{\varepsilon\to 0} \sup_{f\in\mathbb{N}_{\vec{r},d}\left(\vec{\beta},\vec{L}\right)} \boldsymbol{\delta}_\varepsilon^{-\mathfrak{a}} (|\ln(\varepsilon)|)^{\frac{1}{p}} \mathcal{R}_\varepsilon^{(p)}\big[\widehat{f};f\big] \le C.$$

Some remarks are in order.

1^0. Combining the results of Theorems 1 and 2 we conclude that optimally-adaptive estimators under $\mathbb{L}_p$-loss exist over all parameter set $\mathcal{P}^{\rm consist}$ if $p \in \{1,\infty\}$. If $p \in (1,\infty)$ such estimators exist as well except the boundary cases $\kappa(p) = 0$ and $\min_{j=1,\dots,d} r_j = 1$.

2^0. We remark that the upper and lower bound for minimax risk differ each other on the boundary $\kappa(p) = 0$ only by $(|\ln(\varepsilon)|)^{\frac{1}{p}}$-factor. Using $(1,1)$-weak type inequality for strong maximal operator, [2], one can prove adaptive upper bound on the boundary $\min_{j=1,\dots,d} r_j = 1$ containing additional $(|\ln(\varepsilon)|)^{\frac{d-1}{p}}$-factor. Note, nevertheless, that exact asymptotics of minimax risk remains an open problem.

3^0. We obtain full classification of minimax rates over anisotropic Nikolski classes if $p \in \{1,\infty)\}$ and "almost" full one (except the boundaries mentioned above) if $p \in (1,\infty)$. We can assert that $\delta_\varepsilon^{\mathfrak{a}}$ is minimax rate of convergence on $\mathbb{N}_{\vec{r},d}\big(\vec{\beta},\vec{L}\big)$ for any $\vec{\beta} \in (0,\infty)^d, \vec{r} \in (1,\infty]^d$ and $\vec{L} \in (0,\infty)^d$. Indeed, for given $\vec{\beta}$

and $\vec{L}$ one can choose $L_0 = \min_{j=1,\dots d} L_j$ and the number $\ell \in \mathbb{N}^*$ as an any integer strictly larger than $\max_{j=1,\dots d} \beta_j$.

References

[1] Goldenshluger, A. and Lepski, O.V. *On adaptive minimax density estimation on* $\mathbb{R}^d$. Probab. Theory Related Fields, published online 13 july 2013.

[2] de Guzman, M. *Differentiation of Integrals in* R^n. *With appendices by Antonio Córdoba, and Robert Fefferman, and two by Roberto Moriyón.* Lecture Notes in Mathematics, (1975) Vol. 481. Springer-Verlag, Berlin-New York.

[3] Kerkyacharian, G., Lepski, O. and Picard, D.. *Nonlinear estimation in anisotropic multiindex denoising. Sparse case.* Theory Probab. Appl. **52**, (2008) 58–77.

The Multiresolution Statistics for Nonparametric Regression and Inverse Problems

Housen Li

(joint work with Markus Grasmair, Axel Munk)

The problems of nonparametric regression and deconvolution are considered for sub-Gaussian data. A key concept is the *multiresolution (semi-)norm* $\|\cdot\|_{\mathcal{B}}$, defined as

$$\|S_N u\|_{\mathcal{B}} := \sup_{B\in\mathcal{B}} \frac{1}{\sqrt{\#\Gamma_N \cap B}} \left| \sum_{x\in\Gamma_N\cap B} u(x) \right| \quad \text{for } u \in C([0,1]^d),$$

where S_N is the point evaluation on the regular grid Γ_N of $[0,1]^d$ and $\mathcal{B}$ the set of all cubes. This norm has been studied in nonparametric regression problems by many people, such as [1, 2, 3, 4]. In a more general setting of statistical inverse problems, we have studied two multiscale estimates: one in variational form, and the other in constraint form. Both of them take the multiresolution (semi-)norm as data-misfit-measure and the homogeneous Sobolev norm as complexity penalty.

Convergence rates of both approaches are derived in terms of distance function which describes how well the truth can be approximated by functions in the range of the adjoint of forward operator. One crucial point of our results lies in the interpretation of the distance function. This reduces to analyze the asymptotic behavior of the following approximation problem

$$\min\{\|2D^k u - \sum_{x\in\Gamma_N} c_x D^k \varphi_x\|_{L^2} : \|(c_x)_{x\in\Gamma_N}\|_{\mathcal{B}^*} \le t\} \quad \text{as } t \to 0,$$

where φ_x is the solution to

$$(-1)^k \Delta^k \varphi_x = \delta_x - 1.$$

In one dimension case, it relates to the approximation property of B-splines. We have further derived explicit convergence rates, and have shown their optimality and partial adaptivity for certain Sobolev classes. However, in higher dimensions, we do not know the approximation property, especially the behavior of coefficients $(c_x)_{x\in\Gamma_N}$, thus failing to obtain any concrete rate of convergence.

References

[1] A. Nemirovskii, *Nonparametric estimation of smooth regression functions*, Izv. Akad. Nauk. SSR Teckhn. Kibernet. **3** (1985), 50–60 (in Russian), J. Comput. System Sci. **23** (1986), 1–11 (in English).

[2] P. L. Davies and A. Kovac, *Local extremes, runs, strings and multiresolution*, Annals of Statistics, with discussion, **29** (2001), 1–48.

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

[4] K. Frick, A. Munk and H. Sieling, *Multiscale Change-Point Inference*, Journ. Royal Statist. Society, Ser. B, with discussion, To appear.

Geometrizing Statistical Linear Inverse Problems

Tengyuan Liang

(joint work with Tony Cai, Alexander Rakhlin)

Ill-posed inverse problems including high dimensional regression, trace regression, sign vector recovery, orthogonal matrix recovery and permutation matrix recovery pose many challenges for engineers, applied mathematicians and statisticians in the past few years, with techniques such as Dantzig selector, nuclear norm minimization developed to attack each problem. In a recent paper, Chandrasekaran et al. introduced the atomic norm in convex geometry to address a wide class of linear inverse problem simultaneously in the noiseless setting. In our paper, we attack the general linear inverse problems in noisy setting following this line of research. A statistical general linear inverse problem can be formulated in the following way: we want to recover a p dimensional hidden parameter (a vector, matrix or tensor), given n observations of linear transformations of the hidden parameter with independent additive noise. Our research is two folded. Firstly, we show that the local upper bound on rate of convergence of the atomic norm constrained minimization procedure depends on three mathematical terms capturing local convex geometry. In addition, we prove the minimum sample size to ensure the statistical convergence and optimization feasibility of the procedure in terms of dimension, Gaussian width and atomic norm. Secondly, we provide global statistical minimax lower bound for general linear inverse problem, which depends on dimension, sample size and volume ratio driven by the geometry. This is a joint work with Tony Cai and Alexander Rakhlin.

References

[1] T. Cai, T. Liang and A. Rakhlin, *Geometrizing local rate of convergence for statistical linear inverse problems*, manuscript (2013+).

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo and A. S. Willsky, *The convex geometry of linear inverse problems*, Foundations of Computational Mathematics (2010).

[3] P. Bickel, Y. Ritov and A. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, The Annals of Statistics (2009).

[4] G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press (1999).

Optimal Exploration in Multi-Armed Bandit Problems

Robert Nowak
(joint work with Sébastien Bubeck, Kevin Jamieson, and Matt Malloy)

This paper introduces a new algorithm for the *best arm* problem in the stochastic multi-armed bandit (MAB) setting. Consider a MAB with n arms, each with unknown mean payoff $\mu_1, \ldots, \mu_n$ in $[0, 1]$. A sample of the ith arm is an independent realization of a sub-Gaussian random variable with mean μ_i. The goal of the best arm problem is to devise a sampling procedure with a single input δ that, regardless of the values of $\mu_1, \ldots, \mu_n$, finds the arm with the largest mean with probability at least $1 - \delta$. More precisely, best arm procedures must satisfy $\sup_{\mu_1,\ldots,\mu_n} \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$, where i^* is the best arm, $\widehat{i}$ an estimate of the best arm, and the supremum over all sets of means such that there exists a unique best arm.

The best arm problem has a long history dating back to the 1950s [Bechhofer(1958)]. The last decade has seen a flurry of activity providing new upper and lower bounds; see [Even-Dar et al.(2002), Mannor and Tsitsiklis(2004), Jamieson et al.(2013), Karnin et al.(2013)]. The best results show that the best arm can be reliably identified using order $\sum_i \Delta_i^{-2} \log\log \Delta_i^{-2}$ samples, coming within a doubly logarithmic factor of the lower bound of [Mannor and Tsitsiklis(2004)]. Based on the classic work of [Farrell(1964)], we show that in fact that lower bound is not tight, and that the doubly logarithmic factor is necessary. This is a consequence law of the iterated logarithm (LIL), and implies that no procedure can satisfy $\sup_{\Delta_1,\ldots,\Delta_n} \mathbb{P}(\widehat{i} \neq i^*) \leq \delta$ and use fewer than $\sum_i \Delta_i^{-2} \log\log \Delta_i^{-2}$ samples in expectation for all $\Delta_1, \ldots, \Delta_n$.

The LIL also motivates a novel approach to the best arm problem. Specifically, the LIL suggests a natural scaling for confidence bounds on empirical means, and we follow this intuition to develop a new algorithm for the best-arm problem. The algorithm is an Upper Confidence Bound (UCB) procedure [Auer et al.(2002)] based on a finite sample version of the LIL, and so the algorithm is called lil'UCB. By explicitly accounting for the $\log\log$ factor in the confidence bound and using a novel stopping criterion, our analysis of lil'UCB avoids taking naive union bounds over time, as well as the wasteful "doubling trick" employed in algorithms that proceed in epochs [Karnin et al.(2013), Jamieson et al.(2013)]. However, like the algorithm in [Karnin et al.(2013)], lil'UCB is order optimal in terms of the number of samples used to determine the arm with the largest mean.

References

[Bechhofer(1958)] A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

[Even-Dar et al.(2002)] Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.

[Mannor and Tsitsiklis(2004)] The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

[Karnin et al.(2013)] Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[Farrell(1964)] Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, 35(1):pp. 36–72, 1964. ISSN 00034851.

[Jamieson et al.(2013)] On finding the largest mean among many. *arXiv preprint arXiv:1306.3917*, 2013.

[Auer et al.(2002)] Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47 (2-3):235–256, 2002.

Sharp Adaptive Nonparametric Testing for Sobolev Ellipsoids

Michael Nussbaum

(joint work with Pengsheng Ji)

Consider the Gaussian white noise model in sequence space

$$Y_j = f_j + n^{-1/2}\xi_j, \quad j = 1, 2, ...$$

with signal $f = \{f_j\}_{j=1}^{\infty}$ and $\xi_j \sim N(0,1)$ independent. For some $\rho, \beta, M > 0$, consider hypotheses of "no signal" vs. an ellipsoid with l_2-ball removed:

$$H_0 : f = 0 \quad \text{against} \quad H_a : f \in \Sigma(\beta, M) \cap B_\rho,$$

$$B_\rho = \left\{ f \in l_2 : \|f\|_2^2 \geq \rho \right\}, \ \Sigma(\beta, M) = \{ f : \sum_{j=1}^{\infty} j^{2\beta} f_j^2 \leq M \}.$$

Consider α-tests ϕ and their worst case type II error over the alternative:

$$\Psi_n(\phi, \rho, \beta, M) := \sup_{f \in \Sigma(\beta,M) \cap B_\rho} (1 - E_{n,f}\phi). \tag{1}$$

Ingster [8] found the critical rate for $\rho \to 0$, the so-called *separation rate* $\rho_n \asymp n^{-4\beta/(4\beta+1)}$, where a nontrivial type II error behaviour occurs:

$$0 < \liminf \inf_{\phi \ \alpha\text{-test}} \Psi_n \text{ and } \limsup \inf_{\phi \ \alpha\text{-test}} \Psi_n < 1 - \alpha.$$

This rate is known as the *optimal rate for nonparametric testing.* As in nonparametric estimation (cf. Pinsker [13]), the step from optimal minimax rate to optimal constant has been made, with the result by Ermakov [3]:

Suppose $\alpha \in (0,1)$ *and* $\rho_n \sim (cn)^{-4\beta/(4\beta+1)}$ *for some* $c > 0$. *Then*

$$\inf_{\phi \ \alpha\text{-test}} \Psi_n(\phi, \rho, \beta, M) = \Phi(z_\alpha - cM^{-1/4\beta}\eta_\beta) + o(1) \tag{2}$$

where z_α *is the upper* α*-quantile of* $N(0,1)$ *and* $\eta_\beta = (2\beta+1)^{1/2}(4\beta+1)^{-1/2-1/4\beta}$.

We address the question of *sharp minimax adaptive testing,* that is the question of whether this constant can be attained by tests which do not depend on (β, M). For minimax estimation with l_2-loss over ellipsoids $\Sigma(\beta, M)$, cf. Efroimovich and Pinsker [2], Golubev [6], Tsybakov [16]. Adaptation to Pinsker's constant is possible there, without a penalty such as rate loss. For testing, Spokoiny [15] showed that for adaptation to (β, M), there is a rate penalty of order $(\log\log n)^{1/2}$. Essentially this result concerns adaptation to β only; indeed M is irrelevant for the

optimal rate. Moreover, *for adaptation to β only*, Ingster and Suslina [9] obtained a sharp constant, within the $(\log\log n)^{1/2}$ rate loss framework. Adaptation to both parameters (β, M) is an open problem.

We first consider the problem of adaptation to M only, assuming β known.

Theorem 1. *Suppose $c > 0$, $0 < M_1 < M_2 < \infty$ and $\rho_n \sim (cn)^{-4\beta/(4\beta+1)}$. Then there is no test ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$ and both relations*

$$\Psi_n(\phi_n, \rho_n, \beta, M_i) \le \Phi(z_\alpha - cM_i^{-1/4\beta}\eta_\beta) + o(1), \quad i = 1, 2.$$

In view of (2), adaptation to M only is impossible at the separation rate. We now replace the constant c in $\rho_n \sim (cn)^{-4\beta/(4\beta+1)}$ by a sequence $c_n \to \infty$ arbitrarily slowly. Then $\Phi(z_\alpha - c_n M^{-1/4\beta}\eta_\beta) \to 0$, and taking the standard log-asymptotics approach, it turns out that adaptation to Ermakov´s constant is possible.

Theorem 2. *Assume $c_n \to \infty$ and $c_n = o(n^K)$ for every $K > 0$. If $\rho_n = (c_n n)^{-4\beta/(4\beta+1)}$ then there exists a test ϕ_n fulfilling $E_{n,0}\phi_n \le \alpha + o(1)$ and for all $M > 0$*

$$\limsup_n \frac{1}{c_n^2} \log \Psi_n(\phi_n, \rho_n, \beta, M) \le -\frac{M^{-1/2\beta}\eta_\beta^2}{2}.$$

Ermakov [4] showed that the r. h. s. above is also the best achievable for tests possibly depending on M. Hence there is no "penalty for adaptation" here, except that one has to change the optimality criterion. Proofs for this case are in [10].

For the problem of full adaptation to (β, M), we first state a lower asymptotic risk bound for known M and unknown $\beta \in [\beta_1, \beta_2]$, a variation of a result of Ingster and Suslina [9]. Assume that $0 < \beta_1 < \beta_2$ and that $M > 0$ is fixed. Let D be arbitrary and define a radius sequence $\rho_{n,\beta,M}$ by

$$(\rho_{n,\beta,M})^{(4\beta+1)/4\beta} = n^{-1} M^{1/4\beta} \eta_\beta^{-1} \left((2\log\log n)^{1/2} + D \right). \tag{3}$$

Then for any sequence of tests ϕ_n satisfying $E_{n,0}\phi_n \le \alpha + o(1)$

$$\sup_{\beta \in [\beta_1, \beta_2]} \Psi_n(\phi_n, \rho_{n,\beta,M}, \beta, M) \ge (1 - \alpha)\, \Phi(-D) + o(1). \tag{4}$$

Here the test sequences ϕ_n are assumed not to depend on β (but possibly on M); the radius $\rho_{n,\beta,M}$ depends on β and M. The concept of a radius ρ_n varying with β (inside the risk supremum) has been introduced by Spokoiny [15] in the context of rate adaptivity. In the refinement of [9], Ermakov's constant $M^{-1/4\beta}\eta_\beta$ enters the critical radius $\rho_{n,\beta,M}$ as well.

The attainability of the bound (4) is shown in [9] for tests depending on M. We show it for tests not depending on M, establishing adaptivity in (β, M).

Theorem 3. *Assume that $0 < \beta_1 < \beta_2$ and $0 < M_1 < M_2$ are fixed. Let D be arbitrary and let $\rho_{n,\beta,M}$ be the radius sequence in (3). Then there exists a test ϕ_n fulfilling $E_{n,0}\phi_n \le \alpha + o(1)$ and*

$$\sup_{\beta \in [\beta_1, \beta_2], M \in [M_1, M_2]} \Psi_n(\phi_n, \rho_{n,\beta,M}, \beta, M) \le (1 - \alpha)\, \Phi(-D) + o(1).$$

It turns out that there is no additional penalty for M being unknown, and there is no need to consider tail probabilities.

Ingster and Suslina [9] establish their lower bound (4) for l_p- ellipsoids of smoothness r with shrinking l_q- ellipsoids of smoothness s removed, and also Besov classes, but not for sup-norm settings. Lepski and Tsybakov [12] prove a sharp minimax result in testing when the alternative is a Hölder class with a sup-norm ball removed. This represents a testing analog of the minimax estimation result of Korostelev [11] and also a sup-norm analog of Ermakov [3]; for the regression case cf. [5]. When β is given, Dümbgen and Spokoiny [1] establish a sharp adaptivity result with respect to the size parameter M only. The case of unknown (β, M) seems to be an open problem for sup-norm testing; for the estimation case cf. [7]. But in [1] a test is given which is adaptive rate optimal without a $\log\log n$-type penalty. Rohde [14] considers the sup-norm case for regression with nongaussian errors, combining methods of [1] with ideas related to rank tests.

References

[1] L. Dümbgen and V. G. Spokoiny, *Multiscale testing of qualitative hypotheses*, Ann. Statist. **29** (2001), 124–152.

[2] S. Yu. Efroimovich and M. S. Pinsker, *Learning algorithm for nonparametric filtering*, Automation and Remote Control **11** (1984), 1434–1440.

[3] M. S. Ermakov, *Minimax detection of a signal in Gaussian white noise (Russian)*, Teor. Veroyatnost. i Primenen **35** (1990), 704–715; translation in Theory Probab. Appl **35** , 667–679

[4] M. S. Ermakov, *Nonparametric hypothesis testing for small type I and type II error probabilities (Russian)*, Problemy Peredachi Informatsii **44** (2008), no. 2, 54–74; translation in Probl. Inform. Transmission **44**, no. 2, 119–137

[5] G. Gayraud and C. Pouet, C., *Adaptive minimax testing in the discrete regression scheme*, Probab. Theory Related Fields **133** (2005), 531–558.

[6] G. K. Golubev, *Quasilinear estimates for signals in L_2 (Russian)*, Problemy Peredachi Informatsii **26** (1990), no. 1, 19–24; translation in Probl. Inform. Transmission **26** (1990), no. 1, 15–20

[7] G. K. Golubev, O. V. Lepski and B. Levit, *On adaptive estimation for the sup-norm losses*, Math. Methods Statist. **10** (2001), no. 1, 23–37.

[8] Yu. I. Ingster, *Minimax nonparametric detection of signals in white Gaussian noise*. Probl. Inform. Transmission **18** (1982) , no. 2, p. 61

[9] Yu. I. Ingster and I. A. Suslina, *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lecture Notes in Statistics **169**, Springer-Verlag, New York, 2003

[10] P. Ji and M. Nussbaum, *Sharp adaptive nonparametric testing for Sobolev ellipsoids*, arXiv:1210.8162 [math.ST], 2012

[11] A. P. Korostelev, *An asymptotically minimax regression estimator in the uniform norm up to a constant (Russian)*, Teor. Veroyatnost. i Primenen. **38** (1993), 875–882; translation in Theory Probab. Appl. **38**, 737–743

[12] O. V. Lepski and A. B. Tsybakov, *Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point*, Probab. Theory Related Fields **117** (2000), 17–48.

[13] M. S. Pinsker, *Optimal filtration of square-integrable signals in Gaussian noise (Russian)* Problems Inform. Transmission **16** (1980) , no. 2, 52–68

[14] A. Rohde, *Adaptive goodness-of-fit tests based on signed ranks*, Ann. Statist. **36** (2008) 1346–1374.

[15] V. G. Spokoiny, *Adaptive hypothesis testing using wavelets*. Ann. Statist. **24** (1996), 2477–2498

[16] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, NY, 2009.

Nonparametric estimation of the division rate of a Piecewise Deterministic Markov Process: an age model on a tree

Adélaïde Olivier
(joint work with Marc Hoffmann)

We study the evolution of a system of particles. We focus on an individual feature of the particles: their age (but it could be their size, *etc.*). The evolution of the system is driven by two phenomenons. First particles evolve deterministically, here they age. Secondly particles split randomly: a particle of age a splits into two particles of age 0 at a rate $B(a)$. The division rate B is an unknown function we want to estimate.

Recently, Doumic *et al.* in [2] proposed a nonparametric estimation of the division rate B for a size-structured model (where a particle of size x splits in two particles of size $x/2$ at a rate $B(x)$). Their main observation scheme is composed of n cells belonging to the $\lfloor \log_2(n) \rfloor$ first generations. In this work we aim at estimating nonparametrically B observing the evolution of the system until a fixed time T. More precisely we observe the life lengths of the particles that lived before T. Intrinsic difficulties are linked to this observation scheme which is radically different from the previous study [2]. Observing the system between 0 and T introduces first an intricate dependence between datas and also a sampling bias. Intuitively particles that split quickly are more likely to be observed. Bansaye *et al.* [1] proved the first a law of large numbers which makes appear a so called biased density, different from the density associated to B. It enable us to find proper weights to overcome the bias and to recover B through a weighted estimator $\widehat{B}_T$. The weights are estimated since the bias depends on B.

The number of observed particles between 0 and T is random and its expectation when T is large is equivalent to $\exp(\lambda_B T)$ where λ_B is the first eigenvalue of the partial differential equation which describes macroscopically the system, see [3]. We complete the existing law of large numbers giving a rate of convergence for the empirical means. We then exhibit a rate a convergence for our estimator. In squared-loss error over a compact of estimation, $\widehat{B}_T$ centered and renormalized by $\big(\exp(\lambda_B T/2)\big)^{2\beta/(2\beta+1)}$ is tight uniformly on a restricted class of B, if the Hölder-regularity β is large enough. Our estimator is not adaptive with respect to the regularity since the rate is valid for a good choice of window parameter that depends on β. We see here that the magnitude of the speed depends on B, the unknown function we are estimating. Finally we tested numerically our estimator.

References

[1] B. Bansaye, J.-F. Delmas, L. Marsalle, and V. C. Tran. *Limit theorems for Markov processes indexed by continuous time Galton-Watson trees*, The Annals of Applied Probability **21** (2011), 2263–2314.

[2] M. Doumic, M. Hoffmann, N. Krell and L. Robert *Statistical estimation of a growth-fragmentation model observed on a genealogical tree*, Bernoulli, to appear.

[3] B. Perthame *Transport equations arising in biology*, Birckhauser Frontiers in mathematics edition (2007).

On optimal rates for estimation, statistical learning, and online regression

Alexander Rakhlin
(joint work with K. Sridharan and A. Tsybakov)

We consider the problem of regression in three scenarios: (a) regression with random design under the assumption that the model $\mathcal{F}$ is correctly specified, (b) distribution-free statistical learning with respect to a reference class $\mathcal{F}$; and (c) online regression with no assumption on the generative process. To fix notation, let $\mathcal{X}$ be some set and $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[-1, 1]$.

The minimax risk for the first problem (setting (a)) can be written as

$$W_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\|\hat{f} - f\|^2$$

where the infimum is over all estimators based on n i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ distributed as: $X_i \sim P_X$ and $Y_i = f(X_i) + \epsilon_i$, $f \in \mathcal{F}$. Here, ϵ_i is a zero-mean noise, and hence the regression function is $f \in \mathcal{F}$; $\|\cdot\|$ denotes the $L_2(P_X)$-norm.

In the setting of statistical learning, no assumption is placed on the distribution P_{XY} of (X, Y), and the problem is phrased as that of forming a predictor that performs comparably to the best element of the class $\mathcal{F}$:

$$V_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{P_{XY}} \left\{ \mathbb{E}(\hat{f}(X) - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 \right\}$$

Once again, $\hat{f}$ is formed based on an i.i.d. sample of n points from P_{XY}.

Finally, in the online regression scenario, we place no distributional assumptions on the sequence $(x_1, y_1), \ldots, (x_n, y_n)$. To make the problem well-posed, we are asked to predict the sequence sequentially as follows: on round $t = 1, \ldots, n$, we observe $x_t \in \mathcal{X}$, make prediction $\hat{y}_t$ and observe the outcome y_t. The minimax regret in this problem is

$$R_n(\mathcal{F}) = \inf_{\mathcal{A}} \sup_{(x_1,y_1),\ldots,(x_n,y_n)} \left\{ \frac{1}{n} \sum_{t=1}^n (\hat{y}_t - y_t)^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n (f(x_t) - y_t)^2 \right\}$$

where the infimum is taken over all prediction algorithms.

The first problem described above is often studied in the literature on nonparametric estimation, the second falls within the purview of statistical learning theory, and the third is studied within the online learning community. It is recognized that complexity of the class $\mathcal{F}$ plays the key role in determining the minimax behavior; the importance of entropy in the study of estimation goes back to Le Cam, Ibragimov and Khas'minskii, and Birgé. Within the setting of statistical learning the importance of entropy was established in the work of Vapnik and Chervonenkis and in subsequent works on uniform law of large numbers within empirical process theory. The corresponding complexities for online learning have only been found in the past few years. But do these three problems really differ from the minimax point of view?

In this talk, we show that the three problems are, in fact, closely related. In particular, let $\mathcal{N}(\epsilon, \mathcal{F}, d_n)$ be the covering number of $\mathcal{F}$ at scale ϵ with respect to empirical pseudometric d_n. Suppose that $\log \mathcal{N}(\epsilon, \mathcal{F}, d_n) \leq \epsilon^{-p}$. Then, for $p \in (0, 2)$, both $V_n(\mathcal{F})$ and $W_n(\mathcal{F})$ exhibit the rate of $n^{-\frac{2}{2+p}}$. Furthermore, if we place the same assumption on a sequential covering number of $\mathcal{F}$ rather than on $\mathcal{N}(\epsilon, \mathcal{F}, d_n)$, then the same behavior of $n^{-\frac{2}{2+p}}$ can be shown for $R_n(\mathcal{F})$. Both $V_n(\mathcal{F})$ and $R_n(\mathcal{F})$ behave as $n^{-1/p}$ for $p > 2$, signifying a phase transition at $p = 2$, while $W_n(\mathcal{F})$ continues to behave as $n^{-\frac{2}{2+p}}$ for any p. Beyond the equivalence of rates in terms of n, it follows that an algorithm for the third problem can be converted into an algorithm for the first two. Given the new techniques (based on the idea of relaxations) for solving sequential problems such as online regression, there is now hope for developing novel computationally-efficient methods for statistical learning and estimation.

Sparse model selection under heterogeneous noise: exact penalisation and data-driven thresholding

Markus Reiss
(joint work with Laurent Cavalier)

We consider the following sequence space model

$$X_\lambda = f_\lambda + \xi_\lambda, \quad \lambda \in \Lambda, \tag{1}$$

where (f_λ) are the real-valued coefficients of a signal and the noise variables $(\xi_\lambda) \sim N(0, \Sigma)$ have a diagonal covariance matrix $\Sigma = diag(\sigma_\lambda^2)$. Here Λ is a finite, but large index set. This heterogeneous model may appear in several frameworks where the variance is fluctuating, for example in heterogeneous regression, coloured noise, fractional Brownian motion models or especially in statistical inverse problems. For the latter setting the general literature is quite exhaustive, but mostly focusses on specific questions like universal thresholding, asymptotic minimax rates or levelwise thresholding. The aim here is to estimate the unknown parameter vector (f_λ) from the observations (X_λ) under general and unknown sparsity constraints. To this end a penalised empirical risk criterion, based on the so-called risk hull approach, is proposed for general families of possibly data-driven selection rules. This can be viewed as a (data-dependent) model selection procedure and results in a sparse oracle-type inequality.

Model selection is a core problem in statistics. One of the main reference in the field dates back to the information criterion AIC by Akaike, but there is a huge amount of more recent work on this subject, in particular a precise analysis for high-dimensional and sparse data. Model selection is usually linked to the choice of a penalty and its precise choice is the main difficulty in model selection both from a theoretical and a practical perspective. Moreover, there is a close relationship between model selection and the popular thresholding procedure with a false discovery rate (FDR) approach for the threshold choice, cf. [Abramovich *et al* (2006)]. The idea is that the search for a "good penalty" in

model selection is indeed very much related to the choice of a "good threshold" in wavelet procedures. Our main structural assumption is that the parameter vector (f_λ) of interest is sparse, while we do neither know the position nor the number of non-zero entries.

Our goal is to select among a family of models the best possible one, by use of a data-driven selection rule. In particular, one has to deal with the special heterogeneous nature of the observations, which must be reflected by the choice of the penalty. The heterogenous case is much more involved than the direct (homogeneous) model. Indeed, there is no more symmetry inside the stochastic process that one needs to control, since each empirical coefficient has its own variance. The problem and the penalty do not only depend on the number of coefficients that one selects, but also on their position. The penalty is in this sense non-local. We treat the case of general families of data-driven selection rules first and then specify to the full subset selection procedures and the computationally much easier thresholding rules via an FDR-type control. Using our model selection approach, the procedures are almost exact minimax (up to a factor 2 compared to [Golubev (2011)]). Moreover, the procedure is fully adaptive. Indeed, the sparsity index γ_n is unknown and we obtain an explicit penalty, valid in the mathematical proofs and directly applicable in simulations.

The heterogeneity also appears in the minimax lower bounds where the coefficients in the least favourable model will go to the larger variances. In the case of known sparsity γ_n, we consider also a non-adaptive threshold estimator and derive its minimax upper bound. This estimator exactly attains the asymptotic lower bound for typical specifications of the noise levels (σ_λ^2) and is thus exact minimax.

References

[Abramovich *et al* (2006)] Abramovich F., Benjamini Y., Donoho D.L. and Johnstone I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

[Golubev (2011)] Golubev Y. (2011). On oracle inequalities related to data-driven hard thresholding. *Probab. Theory Related Fields* **150**, 435-469.

Simultaneously adaptive estimation for L^2- and L^∞-loss

Johannes Schmidt-Hieber

Consider the Gaussian white noise model $dY_t = f(t)dt + n^{-1/2}dW_t$, $t \in [0,1]$ and let $H(\beta, Q)$ denote the Hölder ball of index β and radius Q. It is well-known that the minimax rate for estimation over $H(\beta, Q)$ with respect to L^2-loss is $n^{-\beta/(2\beta+1)}$. For L^∞-loss, that is, if the loss function is the uniform norm/ supremum norm on $[0,1]$, the minimax rate becomes $(n/\log n)^{-\beta/(2\beta+1)}$. Even if the Hölder index β of the underlying true function is unknown, these rates can be achieved.

Construction of such adaptive estimators for either L^2- *or* L^∞-loss is well-understood. For L^∞-adaptation, hard wavelet thresholding can be used. In practice, this leads, however, to conservative reconstructions in the sense that with high probability no artifacts are included but also rather little of the signal is recovered. In contrast, blockwise thresholding which gives L^2-adaptation with the 'clean' rates $n^{-\beta/(2\beta+1)}$ results in much better reconstructions, but has the drawback that occasionally artificial spikes appear causing the L^2-adaptive estimator to be suboptimal by a $\log(n)$-factor with respect to L^∞-loss (for a precise statement see Theorem 3 in [3]).

A with respect to L^2- and L^∞-loss simultaneously adaptive procedure will inherit the good properties of both methods; because it must also detect small L^2-signal, the procedure will behave similar as blockwise thresholding with (due to the sharp L^∞-rate) less severe artificial spikes.

In the following, let us motivate the construction of the estimator. Decomposing f with respect to the wavelet ψ, we find $f = \sum_{j\geq 0}\sum_k d_{j,k}\psi_{j,k}$, ignoring the scaling coefficients in our heuristic. The transformed observations $Y_{j,k} := \int_0^1 \psi_{j,k}(t)dY_t$ are just the empirical wavelet coefficients $d_{j,k} + n^{-1/2}\epsilon_{j,k}$ with $\epsilon_{j,k} \sim \mathcal{N}(0,1)$, i.i.d. Given $f \in H(\beta, Q)$ implies that there is a constant c, depending on β and Q, with $|d_{j,k}| \leq c2^{-\frac{j}{2}(2\beta+1)}$ for all j,k. Let $J_{n,2}$ and $J_{n,\infty}$ are chosen such that $2^{J_{n,2}} \asymp n^{1/(2\beta+1)}$ and $2^{J_{n,\infty}} \asymp (n/\log n)^{1/(2\beta+1)}$. Now consider the empirical wavelet coefficients $d_{j,k} + n^{-1/2}\epsilon_{j,k}$. All relevant information for adaptive estimation of f with respect to L^2 lies on resolution levels up to $J_{n,2}$. On higher resolutions, the noise dominates the L^2-signal. The same is true for $J_{n,\infty}$ and L^∞-loss. Therefore, we can divide the wavelet decomposition into three parts

$$f = \underbrace{\sum_{j=0}^{J_{n,\infty}}\sum_k d_{j,k}\psi_{j,k}}_{L^2\text{- and } L^\infty\text{-signal}} + \underbrace{\sum_{j=J_{n,\infty}+1}^{J_{n,2}}\sum_k d_{j,k}\psi_{j,k}}_{L^2\text{- but no } L^\infty\text{-signal}} + \underbrace{\sum_{j=J_{n,2}+1}^{\infty}\sum_k d_{j,k}\psi_{j,k}}_{\text{neither } L^2\text{- nor } L^\infty\text{-signal}}.$$

Empirical wavelet coefficients lying in the first part contain signal with respect to both L^2- and L^∞-loss and are driven by noise in the third part of the decomposition. According to the keep-and-kill paradigm in wavelet estimation, we wish to find a procedure keeping empirical wavelet coefficients on low resolutions up to $J_{n,\infty}$ and killing them on resolutions larger than $J_{n,2}$. The critical regime are the resolution level between $J_{n,\infty}$ and $J_{n,2}$ on which empirical wavelet coefficients contain L^2-signal but are pure noise from the L^∞-perspective. Including or excluding the empirical wavelet coefficients from our reconstruction will always lead to suboptimal rates with respect to one of the two losses.

Before studying the general problem, let us consider the case that the smoothness index β and the radius of the Hölder ball Q are known in advance. As mentioned above there is a constant $c = c(\beta, Q)$ with $|d_{j,k}| \leq c2^{-\frac{j}{2}(2\beta+1)}$. Whenever $|Y_{j,k}|$ is larger than $c2^{-\frac{j}{2}(2\beta+1)}$, it is therefore advisable to estimate the wavelet

coefficient by truncation of $Y_{j,k}$, which motivates the estimator

$$\widehat{d}_{j,k} := \operatorname{sign}(Y_{j,k})\big(|Y_{j,k}| \wedge c2^{-\frac{j}{2}(2\beta+1)}\big).$$

It is not difficult to prove that the estimator $\widehat{f} = \sum_{j,k} \widehat{d}_{j,k}\psi_{j,k}$ achieves simultaneously the minimax rates over L^2- and L^∞-loss. Notice, that this estimator does not follow the classical keep-and-kill wavelet thresholding idea. In fact, it acts the opposite way by keeping small empirical wavelet coefficients and truncating (comparably) large ones. Alternatively, one could view the estimator as projection of the empirical wavelet coefficients on the (by the inequalities $|d_{j,k}| \le c2^{-\frac{j}{2}(2\beta+1)}$ slightly enlarged) parameter space.

The question is now, whether we can learn from that for the adaptive problem, where β and Q are unknown and $\widehat{d}_{j,k}$ thus not computable. Estimation of the smoothness index β is very difficult in itself if the true function is 'non-regular'. We can, however, make the following heuristic: If the truth is a regular function, then the bound $c2^{-\frac{j}{2}(2\beta+1)}$ can be estimated rather precisely, whereas if the true function is irregular (roughly speaking, for most points the function lies locally in a smoother space) then simultaneous adaptation for L^2- and L^∞-loss becomes similar to simultaneous adaptation for L^2- and pointwise loss for which no truncation is needed (cf. [1]). In the latter case we may therefore work with a rough upper bound of the truncation level.

Based on the derived heuristics, an estimator can now be constructed which adapts simultaneously over Hölder balls with the clean L^2- and L^∞-rates. Details are in [3].

To conclude, simultaneous estimation with respect to different loss functions enhances and robustifies estimators but might also lead to new procedures and new insights. Many questions remain open (some of them were posed to me by other participants during the workshop). (1) Is it possible to adapt over all L^p-norms simultaneously? (2) Is there an estimator that simultaneously adapts for L^2- and L^∞-loss with the clean rates but now for f *and* its derivatives (cf. also [2]) ? (3) The approach presented above relies crucially on the bounds $|d_{j,k}| \le c2^{-\frac{j}{2}(2\beta+1)}$. Can the ideas be generalized for parameter spaces for which tight inequalities for each wavelet coefficient are unavailable? (4) Is there a more general approach to simultaneous adaptation, in the sense that we can better decide under which conditions simultaneous estimation is possible and how an estimator should be constructed?

References

[1] T. T. Cai, *On block thresholding in wavelet regression: Adaptivity, block size, and threshold level*, Statist. Sinica **12** (2002), 1241–1273.

[2] S. Efromovich, *Simultaneous sharp estimation of functions and their derivatives*, Ann. Statist. **26** (1998), 273–278.

[3] J. Schmidt-Hieber, *Simultaneous L^2- and L^∞-adaptation in nonparametric regression*, preprint, available from `arxiv.org/abs/1303.3118`.

The WiZer, inferred persistence of shape parameters and application to stem cell stress fibre stuctures

Max Sommerfeld

(joint work with Stephan Huckemann, Kwang-Rae Kim, Axel Munk, Florian Rehfeldt, Jochaim Weickert, Carina Wollnik)

We generalize the SiZer of Chaudhuri and Marron (1999, 2000) for the detection of shape parameters of densities on the real line to the case of circular data. The wrapped Gaussian is shown to be the unique choice for this purpose if we require that with increasing levels of smoothing no spurious features are introduced. Based on this we introduce the concept of inferred persistence of shape features and apply this to the analysis of early differentiation in adult human stem cells from their actin-myosin filament structure.

So far, mode and bump hunting has been investigated mainly in the context of (multivariate) density estimation and real line regression.

We are concerned with circular densities for which rigorous inference methods for the number and location of modes have not been provided so far, to the best of our knowledge. Recently, [4] suggested a circular version of the SiZer, however, without providing a circular scale space theory or methods assessing the statistical significance of empirically found modes.

In our work we extend the concept of causality and the SiZer methodology to circular data, inspired on the one hand, by [4] and on the other hand, by a problem arising in studies of early differentiation of human stem cells. We call our estimator based on wrapped Gaussians the WiZer [6]. To this end we

1) propose circular scale space axiomatics,
2) show that under reasonable assumptions the wrapped Gaussian kernel gives the one and only semi-group guaranteeing causality,
3) assess asymptotically the statistical significance of shape features under smoothing with a wrapped Gaussian,
4) and define inferred persistence over smoothing scales of shape features.

References

[1] P. Chaudhuri and J.S. Marron, *Scale space view of curve estimation*, The Annals of Statistics **28** (2000), no. 2, 408–428.

[2] Lutz Dümbgen and Günther Walther, *Multiscale inference about a density*, The Annals of Statistics (2008), 1758–1785.

[3] Adam J Engler, Shamik Sen, H Lee Sweeney, and Dennis E Discher, *Matrix elasticity directs stem cell lineage specification*, Cell **126** (2006), no. 4, 677–689.

[4] María Oliveira, Rosa M Crujeiras, and Alberto Rodríguez-Casal, *Circsizer: an exploratory tool for circular data*, Environmental and Ecological Statistics (2013), 1–17.

[5] J. Schmidt-Hieber, A. Munk, and L. Dümbgen, *Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features*, Ann. Statist. **41** (2013), no. 3, 1299–1328.

[6] S.F. Huckemann, K.-R. Kim. , A. Munk, F. Rehfeld, M. Sommerfeld, J. Weickert, C. Wollnik (2014). *The circular SiZer, inferred persistence of shape parameters and application to stem cell stress fibre structures*, arxiv.org, 1404.3300

[7] A. Zemel, F. Rehfeldt, A. E. X. Brown, D. E. Discher, and S. A. Safran, *Optimal matrix rigidity for stress-fibre polarization in stem cells*, Nat Phys **6** (2010), no. 6, 468–473.

Confidence in credible sets?

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

In Bayesian nonparametrics posterior distributions for functional parameters are often visualized by plotting a center of the posterior distribution, for instance the posterior mean or mode, together with upper and lower bounds indicating a *credible set*, i.e. a set that contains a large fraction of the posterior mass (typically 95%). The credible bounds are intended to visualize the remaining uncertainty in the estimate. In this talk we study the validity of such bounds from a frequentist perspective in the case of priors that are made to adapt to unknown regularity.

It is well known that in infinite-dimensional models Bayesian credible sets are not automatically frequentist confidence sets, in the sense that under the assumption that the data are in actual fact generated by a "true parameter", it is not automatically true that they contain that truth with probability at least the credible level. For a prior of a fixed "regularity level" the (lack of) coverage can be understood in terms of a bias-variance trade-off: Bayesian credible sets typically have *good* frequentist coverage in case of undersmoothing (using a prior that is less regular than the truth), but coverage zero and be far too small in the other case. Simulation studies corroborate theoretical findings, and show that the problem of misleading uncertainty quantification is a very practical one.

The solution to undersmooth the truth, which gives good uncertainty quantification, is unattractive for two reasons. First it leads to a loss in the quality of the reconstruction, e.g. by the posterior mode or mean. Second the true regularity of the functional parameter is never known and hence cannot be used to select a prior that undersmoothes the right regularity. Therefore, in practice it is common to try and "estimate" the regularity from the data and thus to *adapt* the method to the unknown regularity. Bayesian versions of this approach can be implemented using empirical or hierarchical Bayes methods. Empirical Bayes methods estimate the unknown regularity using the marginal likelihood for the data in the Bayesian setup. Hierarchical Bayes methods equip the regularity parameter with a prior and follow a full Bayesian approach.

In this talk we concentrate on the empirical Bayes approach in the context of linear Gaussian inverse problems. The observation is a sequence $X = (X_1, X_2, ...)$ satisfying

$$X_i = \kappa_i \theta_{0,i} + \frac{1}{\sqrt{n}} Z_i, \qquad i = 1, 2, \ldots, \tag{1}$$

where $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ..) \in \ell^2$ is the unknown parameter of interest, the κ_i's are known constants (transforming the truth) and the Z_i are independent, standard normally distributed random variables. The rate of decay of the κ_i's determines

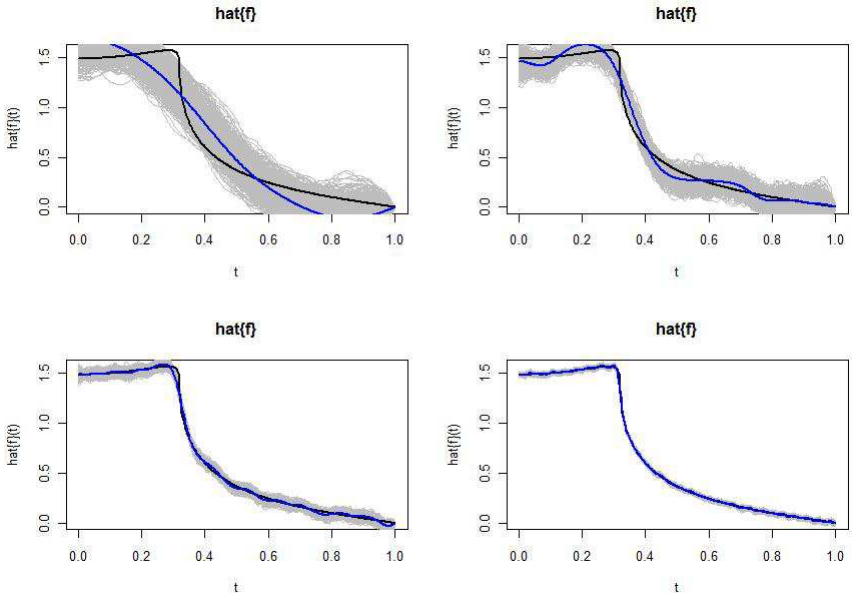

FIGURE 1. Empirical Bayes credible sets. The true function is drawn in black, the posterior mean in blue and the credible set in grey. We have $n = 10^4, 10^6, 10^8$ and 10^{10}, respectively.

the difficulty of the statistical problem of recovering θ_0. We consider the so-called mildly ill-posed case where

$$C^{-2} i^{-2p} \leq \kappa_i^2 \leq C^2 i^{-2p}, \tag{2}$$

for some fixed $p \geq 0$ and $C > 0$. In particular, the choice $p = 0$ corresponds to the ordinary signal-in-white-noise model, whereas $p > 0$ gives a true *inverse problem*.

For $\alpha > 0$ we define a prior measure Π_α for the parameter θ_0 in (1) by

$$\Pi_\alpha = \bigotimes_{i=1}^{\infty} N(0, i^{-1-2\alpha}). \tag{3}$$

The coordinates θ_i are independent under this prior. Since the corresponding coordinates of the data are also independent, the independence is retained in the posterior distribution, which by univariate conjugate Gaussian calculation can be seen to be

$$\Pi_\alpha(\cdot|X) = \bigotimes_{i=1}^{\infty} N\Big(\frac{n\kappa_i^{-1}}{i^{1+2\alpha}\kappa_i^{-2} + n} X_i, \frac{\kappa_i^{-2}}{i^{1+2\alpha}\kappa_i^{-2} + n} \Big). \tag{4}$$

The prior (3) put mass 1 on Sobolev spaces and hyperrectangles of every order strictly smaller than α, and hence expresses a prior belief that the parameter is regular of order (approximately) α. Indeed it can be shown that if the true

parameter θ_0 in (1) belongs to a Sobolev space of order α, then the posterior distribution contracts to the true parameter at the minimax rate $n^{-(1+2\alpha)/(1+2\alpha+2p)}$ for this Sobolev space. A similar result can be obtained for hyperrectangles. On the other hand, if the regularity of the true parameter is different from α, then the contraction can be much slower than the minimax rate.

The suboptimality in the case the true regularity is unknown can be overcome by a data-driven choice of α. The *empirical Bayes* procedure consists in replacing the fixed regularity α in (4) by (for given A, possibly dependent on n)

$$\hat{\alpha}_n = \operatorname*{argmax}_{\alpha\in[0,A]} \ell_n(\alpha), \tag{5}$$

where ℓ_n is the marginal log-likelihood for α in the Bayesian setting: $\theta|\alpha \sim \Pi_\alpha$ and $X|(\theta,\alpha) \sim \otimes_i N(\kappa_i\theta_i, 1/n)$. This is given by

$$\ell_n(\alpha) = -\frac{1}{2}\sum_{i=1}^{\infty}\Big(\log\Big(1+\frac{n}{i^{1+2\alpha}\kappa_i^{-2}}\Big) - \frac{n^2}{i^{1+2\alpha}\kappa_i^{-2}+n}X_i^2\Big). \tag{6}$$

If there exist multiple maxima, any one of them can be chosen.

The *empirical Bayes posterior* is defined as the random measure $\Pi_{\hat{\alpha}_n}(\cdot|X)$ obtained by substituting $\hat{\alpha}_n$ for α in the posterior distribution (4), i.e.

$$\Pi_{\hat{\alpha}_n}(\cdot|X) = \Pi_\alpha(\cdot|X)\Big|_{\alpha=\hat{\alpha}_n}$$

This adapted posterior can be shown to contract to the true parameter at the (near) minimax rate within the setting of Sobolev balls and hyperrectangles.

For fixed $\alpha > 0$, let $\hat{\theta}_{n,\alpha}$ be the posterior mean corresponding to the prior Π_α (see (4)). The centered posterior is a Gaussian measure that does not depend on the data and hence for $\gamma \in (0,1)$ there exists a deterministic radius $r_{n,\gamma}(\alpha)$ such that the ball around the posterior mean with this radius receives a fraction $1-\gamma$ of the posterior mass, i.e. for $\alpha > 0$,

$$\Pi_\alpha\big(\theta : \|\theta - \hat{\theta}_{n,\alpha}\| \le r_{n,\gamma}(\alpha)|X\big) = 1-\gamma. \tag{7}$$

In the exceptional case that $\alpha = 0$ we define the radius to be infinite. The empirical Bayes credible sets that we consider in this paper are the sets obtained by replacing the fixed regularity α by the data-driven choice $\hat{\alpha}_n$. Here we introduce some more flexibility by allowing the possibility of blowing up the balls by a factor L. For $L > 0$ we define

$$\hat{C}_n(L) = \{\theta \in \ell^2 : \|\theta - \hat{\theta}_{n,\hat{\alpha}_n}\| \le L r_{n,\gamma}(\hat{\alpha}_n)\}. \tag{8}$$

By construction $\Pi_{\hat{\alpha}_n}(\hat{C}_n(L)|X) \ge 1-\gamma$ iff $L \ge 1$.

We are interested in the performance of the random sets $\hat{C}_n(L)$ as frequentist confidence sets. Ideally we would like them to be *honest* in the sense that

$$\inf_{\theta_0\in\Theta_0} P_{\theta_0}\big(\theta_0 \in \hat{C}_n(L)\big) \ge 1-\gamma,$$

for a model Θ_0 that contains all parameters deemed possible. In particular, this model should contain parameters of all regularity levels. At the same time we would like the sets to be *adaptive*, in the sense that the radius of $\hat{C}_n(L)$ is (nearly)

bounded by the optimal rate for a model of a given regularity level, whenever θ_0 belongs to this model. It is well known that this is too much to ask, as there do not exist confidence sets with this property, Bayesian or non-Bayesian. For the present procedure we can explicitly exhibit examples of "inconvenient truths" that are not covered at all.

Theorem 3. *For given positive integers n_j with $n_1 \geq 2$ and $n_j \geq n_{j-1}^4$ for every j, $\beta > 0$ and $M > 0$, define $\theta_0 = (\theta_{0,1}, \theta_{0,2}, ...)$ by*

$$\theta_{0,i}^2 = \begin{cases} 2^{-1-2\beta} M n_j^{-\frac{1+2\beta}{1+2\beta+2p}}, & \text{if } n_j^{\frac{1}{1+2\beta+2p}} \leq i < 2n_j^{\frac{1}{1+2\beta+2p}}, \qquad j = 1, 2, \ldots, \\ 0, & \text{otherwise} \end{cases}$$

Then the constant $M > 0$ can be chosen such that $P_{\theta_0}(\theta_0 \in \hat{C}_{n_j}(L_n)) \to 0$ as $j \to \infty$ for every $L_n \ll n^{(1/2+p)/((1+2\beta+2p)(2+2\beta+2p))}$.

By construction the (fixed) parameter θ_0 defined in Theorem 3 belongs to the hyperrectangle $\Theta^\beta(M)$, and in this sense is a "good", because "smooth" truth. However, it is an inconvenient truth, as it tricks the empirical Bayes procedure, making this choose the "wrong" regularity α, for which the corresponding credible set does not cover θ_0.

Intuitively it is not surprising that such bad behaviour occurs, as nonparametric credible or confidence sets always necessarily extrapolate into aspects of the truth that are not visible in the data. Honest uncertainty quantification is only possible by a-priori assumptions on those latter aspects. In the context of regularity this may be achieved by "undersmoothing", for instance by using a prior of fixed regularity smaller than the true regularity. Alternatively we may change the notion of regularity and strive for honesty over different models. In the latter

spirit we shall show that the empirical Bayes credible sets $\hat{C}_n(L)$ are honest over classes of "polished" truths.

Definition 2. *A parameter $\theta \in \ell^2$ satisfies the* polished tail condition *if, for fixed positive constants L_0, N_0 and $\rho \geq 2$,*

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L_0 \sum_{i=N}^{\rho N} \theta_i^2, \qquad \forall N \geq N_0. \tag{9}$$

We denote by $\Theta_{pt}(L_0, N_0, \rho)$ the set of all polished tail sequences $\theta \in \ell^2$ for the given constants L_0, N_0 and ρ.

It can be shown that the set of all polished tail sequences is nearly equal to all of ℓ_2 in a topological sense; that the statistical problem does not become easier if one knows that the true parameter is polished tail; and that almost every parameter generated from one of the priors Π_α is polished tail. Thus, in a sense, assuming that the true parameter is polished tail is natural.

Under this assumption the adaptive credible sets perform reasonably well.

Theorem 4. *For any A, L_0, N_0 there exists a constant L such that*

$$\inf_{\theta_0 \in \Theta_{pt}(L_0)} P_{\theta_0}\big(\theta_0 \in \hat{C}_n(L)\big) \to 1. \tag{10}$$

Furthermore, for $A = A_n \leq \sqrt{\log n}/(4\sqrt{\log \rho \vee e})$ *this is true with a slowly varying sequence* ($L := L_n \leq C(3\rho^{3(1+2p)})^{A_n}$ *works).*

Impact of Regularization on Spectral Clustering

Bin Yu
(joint work with Antony Joseph)

The problem of identifying communities, or clusters, in large networks is an important contemporary problem in statistics. Spectral clustering is one of the more popular techniques for such purposes, chiefly due to its computational advantage and generality of application. The algorithm's generality arises from the fact that it is not tied to any modeling assumptions on the data, but is rooted in intuitive measures of community structure such as *sparsest cut* based measures [8], [16], [10], [13]. Other examples of applications of spectral clustering include manifold learning [2], image segmentation [16], and text mining [6].

The canonical nature of spectral clustering also generates interest in variants of the technique. Here, we attempt to better understand the impact of regularized forms of spectral clustering for community detection in networks. In particular, we focus on the regularized spectral clustering (RSC) procedure proposed in [1]. Their empirical findings demonstrates that the performance of the RSC algorithm, in terms of obtaining the correct clusters, is significantly better for certain values of the regularization parameter. An alternative form of regularization was studied in [5], and [14].

We attempt to provide a theoretical understanding for the regularization in the RSC algorithm. Our analysis focuses on the Stochastic Block Model (SBM) and an extension of this model. We also address the practical issue of the choice of regularization parameter.

Our results involves understanding the interplay, as a function of the regularization parameter, between the *eigen gap* and the concentration of the sample Laplacian. Assuming that there are K clusters, the eigen gap refers to the gap between the K-th smallest eigenvalue and the remaining eigenvalues. An adequate gap ensures that the sample eigenvectors can be estimated well, [18], [13], [10], which leads to good cluster recovery.

The adequacy of an eigen gap for cluster recovery is in turn determined by the concentration of the sample Laplacian. In particular, a consequence of the Davis-Kahan theorem [3] is that if the spectral norm of the difference of the sample and population Laplacians is small compared to the eigen gap then the top K eigenvector can be estimated well. Denoting τ as the regularization parameter, previous theoretical analyses of regularization [5], [15], provided high-probability bounds on this spectral norm. Denoting d_{min} as the minimum expected degree of the graph, these bounds have a $1/\sqrt{\tau + d_{min}}$ dependence on τ. In contrast, our high probability bounds behave like $1/\tau$ as τ varies. The end result is that we show that one can get a good understanding of the impact of regularization by

understanding the situation where τ goes to infinity. This also explains empirical observations in [1], [14] where it was seen that performance of regularized spectral clustering does not change for τ beyond a certain value. Below are the three main contributions of the talk.

We attempt to understand regularization for the stochastic block model. In particular, for a graph with n nodes, previous theoretical analyses for spectral clustering, under the SBM and its extensions, [15],[5], [17], [7] assumed that the minimum degree of the graph scales at least by a polynomial power of $\log n$. Even when this assumption is satisfied, the dependence on the minimum degree is highly restrictive when it comes to making inferences about cluster recovery. Our analysis provides cluster recovery results that potentially do not depend on the above mentioned constraint on the minimum degree. As an example, for an SBM with two blocks (clusters), our results depend on the average degree, as opposed to the minimum degree.

Further, we demonstrate that regularization has the potential of addressing a situation, often encountered in practice, where not all nodes belong to well-defined clusters. Without regularization, these nodes would hamper with the clustering of the remaining nodes in the following way: In order for spectral clustering to work, the top eigenvectors - that is, the eigenvectors corresponding to the largest eigenvalues of the Laplacian - need to be able to discriminate between the clusters. Due to the effect of nodes that do not belong to well-defined clusters, these top eigenvectors do not necessarily discriminate between the clusters with ordinary spectral clustering. With a proper choice of regularization parameter, we show that this problem can be rectified. We also demonstrate this on simulated and real datasets.

Although our theoretical results deal with the 'large' τ case, it is observed empirically that moderate values of τ may produce slightly better clustering performance. Consequently, we also propose a data dependent procedure for choosing the regularization parameter. The procedure works by providing estimates of the Davis-Kahan bounds over a grid of values of τ and then choosing the τ that minimizes these estimates. We demonstrate that this works well through simulations and on a real data set.

References

[1] A.A. Amini, A. Chen, P.J. Bickel, and E. Levina. Fitting community models to large sparse networks. *Ann. Statist*, 41(4):2097–2122, 2013.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer, 1997.

[4] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[5] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23.

[6] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. seventh ACM SIGKDD inter. conf. on Know. disc. and data mining*, pages 269–274. ACM, 2001.
[7] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013.
[8] Lars Hagen and Andrew B Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11(9):1074–1085, 1992.
[9] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.
[10] Tsz Chiu Kwok, Lap Chi Lau, Yin Tat Lee, Shayan Oveis Gharan, and Luca Trevisan. Improved cheeger's inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. *arXiv preprint arXiv:1301.5584*, 2013.
[11] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.
[12] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
[13] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.
[14] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *arXiv preprint arXiv:1309.4111*, 2013.
[15] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.
[16] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pat. Analysis and Mach. Intel.*, 22(8):888–905, 2000.
[17] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.
[18] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Structured Matrix Completion

Anru Zhang
(joint work with Tianxi Cai, T. Tony Cai)

Matrix completion has attracted significant recent attention in many fields including statistics, applied mathematics and electrical engineering. Current literature on matrix completion focuses primarily on independent sampling models under which the individual observed entries are sampled independently. Motivated by applications in genomic data integration, we propose a new framework of structured matrix completion (SMC) to treat structured missingness by design. Specifically, our proposed method aims at efficient matrix recovery when a subset of the rows and columns of an approximately low-rank matrix are observed. We provide theoretical justification for the proposed SMC method and derive lower bound for the estimation errors, which together establish the optimal rate of recovery over certain classes of approximately low-rank matrices. Simulation studies show that the method performs well in finite sample under a variety of configurations.

References

[1] T. Cai, T. T. Cai, A. Zhang, *Structured matrix completion with applications in genomic data integration*, technical report, (2014).

Asymptotic normality and optimalities in estimation of large Gaussian graphical model

Harrison Zhou

(joint work with Zhao Ren, Tingni Sun, Cun-Hui Zhang)

Gaussian graphical model, a powerful tool for investigating the relationship among a large number of random variables in a complex system, is used in a wide range of scientific applications. A central question for Gaussian graphical model is to recover the structure of an undirected Gaussian graph. Let $G = (V, E)$ be an undirected graph representing the conditional dependence relationship between components of a random vector $Z = (Z_1, \ldots, Z_p)^T$ as follows. The vertex set $V = \{V_1, \ldots, V_p\}$ represents the components of Z. The edge set E consists of pairs (i, j) indicating the conditional dependence between Z_i and Z_j given all other components. In applications, the following question is fundamental: Is there an edge between V_i and V_j? It is well known that recovering the structure of an undirected Gaussian graph $G = (V, E)$ is equivalent to recovering the support of the population precision matrix of the data in the Gaussian graphical model. Let

$$Z = (Z_1, Z_2, \ldots, Z_p)^T \sim \mathcal{N}(\mu, \Sigma),$$

where $\Sigma = (\sigma_{ij})$ is the population covariance matrix. The precision matrix, denoted by $\Omega = (\omega_{ij})$, is defined as the inverse of covariance matrix, $\Omega = \Sigma^{-1}$. There is an edge between V_i and V_j, i.e., $(i, j) \in E$, if and only if $\omega_{ij} \neq 0$. Consequently, the support recovery of the precision matrix Ω yields the recovery of the structure of the graph G.

Suppose n i.i.d. p-variate random vectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are observed from the same distribution as Z, i.e. the Gaussian $\mathcal{N}\left(\mu, \Omega^{-1}\right)$. Assume without loss of generality that $\mu = 0$. In this paper, we address the following two fundamental questions: When is it possible to make statistical inference for each individual entry of a precision matrix Ω at the parametric $\sqrt{n}$ rate? When and in what sense is it possible to recover the support of Ω in the presence of some small nonzero $|\omega_{ij}|$?

The problems of estimating a large sparse precision matrix and recovering its support have drawn considerable recent attention. In spite of an extensive literature on the topic, it is still largely unknown the fundamental limit of support recovery in the Gaussian graphical model, let alone an adaptive procedure to achieve the limit.

This paper makes important advancements in the understanding of statistical inference of low-dimensional parameters in the Gaussian graphical model in the following ways. Let s be the maximum degree of the graph or a certain more relaxed capped-ℓ_1 measure of the complexity of the precision matrix. We prove

that the estimation of each ω_{ij} at the parametric $\sqrt{n}$ convergence rate requires the sparsity condition $s \leq O(1)n^{1/2}/\log p$ or equivalently a sample size of order $(s \log p)^2$. We propose an adaptive estimator of individual ω_{ij} and prove its asymptotic normality and efficiency when $n \gg (s \log p)^2$. Moreover, we prove that the proposed estimator achieves the optimal convergence rate when the sparsity condition is relaxed to $s \leq c_0 n/\log p$ for a certain positive constant c_0. The efficient estimator of the individual ω_{ij} is then used to construct fully data driven procedures to recover the support of Ω and to make statistical inference about latent variables in the graphical model.

The methodology we are proposing is a novel regression approach. For any index subset A of $\{1, 2, \ldots, p\}$ and a vector Z of length p, we use Z_A to denote a vector of length $|A|$ with elements indexed by A. Similarly for a matrix U and two index subsets A and B of $\{1, 2, \ldots, p\}$ we can define a submatrix $U_{A,B}$ of size $|A| \times |B|$ with rows and columns of U indexed by A and B respectively. Consider $A = \{i, j\}$, for example, $i = 1$ and $j = 2$, then $Z_A = (Z_1, Z_2)^T$ and $\Omega_{A,A} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}$. It is well known that

$$Z_A | Z_{A^c} = \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c} Z_{A^c}, \Omega_{A,A}^{-1}\right).$$

This observation motivates us to consider regression with two response variables above. The noise level $\Omega_{A,A}^{-1}$ has only three parameters. When Ω is sufficiently sparse, a penalized regression approach is proposed to obtain an asymptotically efficient estimation of ω_{ij}, i.e., the estimator is asymptotically normal and the variance matches that of the maximum likelihood estimator in the classical setting where the dimension p is a fixed constant. Consider the class of parameter spaces modeling sparse precision matrices with at most $k_{n,p}$ off-diagonal nonzero elements in each column,

$$(1)\quad \mathcal{G}_0(M, k_{n,p}) = \left\{ \begin{array}{c} \Omega = (\omega_{ij})_{1 \leq i,j \leq p} : \max_{1 \leq j \leq p} \sum_{i \neq j} 1\{\omega_{ij} \neq 0\} \leq k_{n,p}, \\ \text{and } 1/M \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M. \end{array} \right\},$$

where $1\{\cdot\}$ is the indicator function and M is some constant greater than 1. The following theorem shows that a necessary and sufficient condition to obtain a $\sqrt{n}-$consistent estimation of ω_{ij} is $k_{n,p} = O\left(\frac{\sqrt{n}}{\log p}\right)$, and when $k_{n,p} = o\left(\frac{\sqrt{n}}{\log p}\right)$ the procedure to be proposed is asymptotically efficient.

Theorem. Let $X^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$, $i = 1, 2, \ldots, n$. Assume that $k_{n,p} \leq c_0 n/\log p$ with a sufficiently small constant $c_0 > 0$ and $p \geq k_{n,p}^{\nu}$ with some $\nu > 2$. We have the following probablistic results,

(i): There exists a constant $\epsilon_0 > 0$ such that

$$\inf_{i,j} \inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M, k_{n,p})} \mathbb{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| \geq \epsilon_0 \max\left\{n^{-1} k_{n,p} \log p, n^{-1/2}\right\}\right\} \geq \epsilon_0.$$

(ii): The estimator $\hat{\omega}_{ij}$ proposed is rate optimal in the sense of

$$\max_{i,j} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{P}\Big\{ |\hat{\omega}_{ij} - \omega_{ij}| \geq M \max\left\{n^{-1}k_{n,p}\log p, n^{-1/2}\right\}\Big\} \to 0,$$

as $(M,n) \to (\infty,\infty)$. Furthermore, the estimator $\hat{\omega}_{ij}$ is asymptotically efficient when $k_{n,p} = o\left(\frac{\sqrt{n}}{\log p}\right)$, i.e., with $F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2$,

$$\sqrt{nF_{ij}}\left(\hat{\omega}_{ij} - \omega_{ij}\right) \xrightarrow{D} \mathcal{N}(0,1). \tag{2}$$

Moreover, the minimax risk of estimating ω_{ij} over the class $\mathcal{G}_0(k, M_{n,p})$ satisfies, provided $n = O\left(p^{\xi}\right)$ with some $\xi > 0$,

$$\inf_{\hat{\omega}_{ij}} \sup_{\mathcal{G}_0(M,k_{n,p})} \mathbb{E}\,|\hat{\omega}_{ij} - \omega_{ij}| \asymp \max\left\{k_{n,p}\frac{\log p}{n}, \sqrt{\frac{1}{n}}\right\}. \tag{3}$$

The lower bound is established through Le Cam's Lemma and a novel construction of a subset of sparse precision matrices. An important implication of the lower bound is that the difficulty of support recovery for sparse precision matrix is different from that for sparse covariance matrix when $k_{n,p} \gg \left(\frac{\sqrt{n}}{\log p}\right)$, and when $k_{n,p} \ll \left(\frac{\sqrt{n}}{\log p}\right)$ the difficulty of support recovery for sparse precision matrix is just the same as that for sparse covariance matrix.

It is worthwhile to point out that the asymptotic efficiency result is obtained without the need to assume the irrepresentable condition or the l_1 constraint of the precision matrix which are commonly required in literature. An immediate consequence of the asymptotic normality result (2) is to test individually whether there is an edge between V_i and V_j in the set E, i.e., the hypotheses $\omega_{ij} = 0$. The result is applied to do adaptive support recovery optimally. In addition, we can strengthen other results in literature under weaker assumptions, and the procedures are adaptive, including adaptive rate-optimal estimation of the precision matrix under various matrix l_q norms, and an extension of our framework for inference and estimation to a class of latent variable graphical models.

Reporters: Axel Munk, Alexandre Tsybakov

Participants

Prof. Dr. Lucien Birgé
Laboratoire de Probabilités-Tour 56
Université P. et M. Curie
4, Place Jussieu
75252 Paris Cedex 05
FRANCE

Prof. Dr. Lawrence D. Brown
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

Victor-Emmanuel Brunel
École Nationale de la Statistique
et de l'Adm. Economique
ENSAE
3, avenue Pierre Larousse
92245 Malakoff Cedex
FRANCE

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

Prof. Dr. Florentina Bunea
Department of Statistical Science
Cornell University
Comstock Hall
Ithaca, NY 14853-2601
UNITED STATES

Prof. Dr. Cristina Butucea
Lab. d'Analyse et de Mathématiques
Appl.
UFR Mathématiques
Université Paris-Est Marne-la-Vallée
5, Bd. Descartes, Champs sur Marne
77454 Marne-la-Vallée Cedex 2
FRANCE

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

Prof. Dr. Rainer Dahlhaus
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Prof. Dr. Arnak Dalalyan
ENSAE / CREST
École Nationale de la Statistique et de
l'Administration Économique
3, avenue Pierre Larousse
92245 Malakoff Cedex
FRANCE

Prof. Dr. Holger Dette
Fakultät für Mathematik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Manuel Diehn
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen
GERMANY

Prof. Dr. Lutz Dümbgen
Institut für mathematische Statistik
und Versicherungslehre
Universität Bern
Alpeneggstr. 22
3012 Bern
SWITZERLAND

Dr. Eric Gautier
Centre de Recherches en Economie et
Statistiques
15 Boulevard Gabriel Peri
92245 Malakoff Cedex
FRANCE

Prof. Dr. Alexander Goldenshluger
Department of Statistics
University of Haifa
Haifa 31905
ISRAEL

Prof. Dr. Laszlo Györfi
Department of Computer Science and
Information Theory
Budapest University of Techn. &
Economics
Stoczek u. 2
1521 Budapest
HUNGARY

Prof. Dr. Marc Hoffmann
École Nationale de la Statistique
e de l'Adm. Economique
ENSAE
3, avenue Pierre-Larousse
92245 Malakoff
FRANCE

Prof. Dr. Thorsten Hohage
Institut f. Numerische & Angew.
Mathematik
Universität Göttingen
Lotzestr. 16-18
37083 Göttingen
GERMANY

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

Dr. Olga Klopp
MODAL'X
Université Paris Ouest Nanterre La
Défense
200 avenue de la République
92001 Nanterre Cedex
FRANCE

Dr. Guillaume Lecué
Centre de Mathématiques Appliquées
CMAP - UMR 7641
Ecole Polytechnique
Route de Saclay
91120 Palaiseau Cedex
FRANCE

Prof. Dr. Oleg Lepski
Centre de Mathématiques et
d'Informatique
Université de Provence
39, Rue Joliot-Curie
13453 Marseille Cedex 13
FRANCE

Housen Li
Max-Planck-Institut
für Biophysikalische Chemie
Am Fassberg 11
37077 Göttingen
GERMANY

Tengyuan Liang
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

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

Prof. Dr. Enno Mammen
Abteilung f. Volkswirtschaftslehre
Universität Mannheim
L 7, 3-5
68131 Mannheim
GERMANY

Dr. Katia Meziani
CEREMADE
Université Paris Dauphine
Place du Marechal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

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

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

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

Adélaide Olivier
École Nationale de la Statistique
et de l'Administration Economique
ENSAE CREST - Laboratoire de
Statistique
3, avenue Pierre-Larousse
92245 Malakoff
FRANCE

Florian Pein
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen
GERMANY

Prof. Dr. Marianna Pensky
Department of Mathematics
University of Central Florida
Orlando, FL 32816-1364
UNITED STATES

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

Prof. Dr. Alexander Rakhlin
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

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

Prof. Dr. Angelika Rohde
Ruhr-Universität Bochum
Fakultät für Mathematik
Lehrstuhl für Stochastik
44780 Bochum
GERMANY

Prof. Dr. Judith Rousseau
CEREMADE
Université Paris Dauphine
Place du Marechal De Lattre de Tassigny
75016 Paris Cedex
FRANCE

Till Sabel
Institut f. Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen
GERMANY

Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 OWB
UNITED KINGDOM

Dr. Johannes Schmidt-Hieber
École Nationale de la Statistique
e de l'Adm. Economique
ENSAE
3, avenue Pierre-Larousse
92245 Malakoff
FRANCE

Prof. Dr. Catia Scricciolo
Department of Decision Sciences
"L. Bocconi" University
Via G. Röntgen 1
20136 Milano
ITALY

Max Sommerfeld
Statistical and Applied Mathematical
Science Institute (SAMSI)
19 T.W. Alexander Drive
P.O. Box 14006
Research Triangle Park, NC 27709-4006
UNITED STATES

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

Prof. Dr. Alexandre B. Tsybakov
CREST
Timbre J 340
3, av. P. Larousse
92240 Malakoff Cedex
FRANCE

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich
HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

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

Prof. Dr. Dan Yang
Department of Statistics & Biostatistics
Rutgers University
453 Hill Center
110 Frelinghuysen Road
Piscataway, NJ 08854-8019
UNITED STATES

Prof. Dr. Bin Yu
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Anru Zhang
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

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

Prof. Dr. Linda Zhao
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104-6340
UNITED STATES

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