Mathematisches Forschungsinstitut Oberwolfach

Report No. 16/2010

DOI: 10.4171/OWR/2010/16

Modern Nonparametric Statistics: Going Beyond Asymptotic Minimax

Organised by
Lucien Birgé, Paris
Iain M. Johnstone, Stanford
Vladimir Spokoiny, Berlin

March 28th – April 3rd, 2010

Abstract. During the years 1975 - 1990 a major emphasis in nonparametric estimation was put on computing the asymptotic minimax risk for many classes of functions. Modern statistical practice indicates some serious limitations of the asymptotic minimax approach and calls for some new ideas and methods which can cope with the numerous challenges brought to statisticians by modern sets of data.

Mathematics Subject Classification (2000): 62Gxx.

Introduction by the Organisers

The workshop took place during the period *March 28 - April 2* and, as usual, talks were planned from Monday morning to Friday morning (most participants leaving on Friday afternoon) with a break on Wednesday afternoon for the traditional walk to Saint-Roman.

There were finally 48 participants, due to some late cancellations. Unfortunately, Iain Johnstone could not attend the meeting since he had a very important committment in the US with the NSF during that week. However, he could participate quite actively in the organization up to the last minute since we, organizers, had the opportunity to meet together during a previous workshop and also exchange extensively by e-mail through which the list of participants and talks and all final details were set up. Therefore we were really three organizers and the success of the meeting should be put on the three of us. Actually, the list of speakers and the schedule of the talks were ready before our arrival and only minor changes

were made during that week. This precise schedule can be found at the end of our report.

During the years 1975 - 1990 (roughly speaking) a major emphasis in nonparametric estimation was put on computing the (possibly asymptotic) minimax risk for many classes of functions, starting from the simplest Hölder classes to the more sophisticated Besov balls in the beginning of the 90's. It was clear, at that time, that this minimax point of view was quite pessimistic, since it was directed towards the worse case and also unrealistic, since one never knows to which smoothness class (or other specific class) the true parameter does belong. Nevertheless, this approach allowed to design useful estimators, which could be more or less practically calibrated (by cross-validation for instance) and provided some benchmarks for the performance of a given method.

Then, by the beginning of the 90's (approximately), started an important movement towards what is now called *adaptation*, either to some smoothness class or to the specific function that was to be estimated. This was made via different tools like Lepski's method, the use of localized basis and thresholding, model selection ...

More recently, many new methods (aggregation of estimators, Lasso, etc.) appeared in order to cope with the numerous challenges brought to statisticians by modern sets of data and the huge progress of computing : huge data sets or situations where the number of unknown parameters is much larger than the number of data, together with some sparsity assumption. This also coincides with an important renewal of Bayesian methods due to much better and powerful computing facilities.

Workshop organization

In view of the importance of the numerous new techniques that are presently studied and used to solve the challenges offered by the modern sets of data, we decided that the main purpose of the workshop would be to expose many young researchers to those new techniques. We invited a number of confirmed specialists and experts together with younger professionals, PhD. students, postdocs, new assistant professors, in order to get a mix of generations and experiences. We also selected 5 senior professors to give longer talks (one hour and a half, one each morning) in order to develop their subject. These persons were especially asked, several months before the workshop, to deliver these special conferences.

We also spent a lot of time and discussion in order to select the talks among the proposals by the participants in order to keep a maximal coherence between the subjects and keep the level as high as possible, finally limiting the number of talks to 24, including the five major ones mentioned above, and avoiding the multiplication of short talks. All normal talks were of 45mn, with the exception of the last morning since it was asked to us by the MFO organization to shorten the session for an early lunch (apparently for Easter vacation).

It was also an occasion for us to invite an unusually large number of participants (mostly young researchers and some more senior French) that visited the MFO for

the first time, which gave them an occasion to discover this very nice place, the wonderful library, the numerous working facilities and the excellent MFO organization (as usual).

We tried, as much as possible, to organize our 8 sessions around themes like Model Selection, Adaptive Density Estimation, High-dimensional Data and Sparsity, Statistics for Processes, Nonparametric Bayesian Methods, with also some talks by young and promising researchers which were given exactly the same time as the more senior ones.

Workshop: Modern Nonparametric Statistics: Going Beyond Asymptotic Minimax

Table of Contents

Abstracts

Adapting to inhomogeneous and anisotropic smoothness via dyadic partition selection

Nathalie Akakpo

Our research work is devoted to adaptive functional estimation by selection of a best partition into dyadic intervals, cubes or rectangles. Our estimation procedure can be described as follows. Assume that you want to estimate some function s, which is square integrable over $[0,1]^d$ ($d \in \mathbb{N}^\star$) on the basis of the observation of n independent random variables. For instance, s may be the marginal density of an i.i.d. sample or a regression function. We first give ourselves a collection of partitions into dyadic rectangles with prescribed minimal sidelength. On each partition, we define a piecewise polynomial estimator obtained by minimization of an adequate least-squares type criterion. Then, we select from the data the best estimator in that collection by using a penalized criterion. Such a procedure thus generalizes the one introduced by Donoho [4] in a regression framework.

From a theoretical point of view, we obtain two kinds of results. On the one hand, our estimator satisfies in many frameworks non-asymptotic oracle-type inequalities, up to a factor that does not depend on the size n of the sample. On the other hand, it reaches the minimax risk, still up to a factor that does not depend on n, over classes of functions that may be of inhomogeneous and anisotropic smoothness. Such a property – that, up to our knowledge, has never been shown for any estimator – heavily relies on approximation results in the spirit of the paper [3] by DeVore and Yu that we prove in [1]. Moreover, our estimator can be implemented with a complexity which is linear in the sample size.

In this talk, we present that procedure in the framework of conditional density estimation, which is part of the joint work [2] with Claire Lacour (Université Paris Sud).

References

[1] N. Akakpo, *Estimation adaptative par sélection de partitions en rectangles dyadiques*, Ph.D. Thesis, Université Paris Sud, Orsay (2009).

[2] N. Akakpo and C. Lacour, *Inhomogeneous and anisotropic conditional density estimation from dependent data*, Working paper (2010).

[3] R.A. DeVore and X.M. Yu, *Degree of adaptive approximation*, Math. Comp. **55** (1990), 625–635.

[4] D. L. Donoho, *CART and best-ortho-basis: a connection*, Ann. Statist., **25**, 5 (1997), 1870–1911.

Data-driven penalties for linear estimators selection

Sylvain Arlot
(joint work with Francis Bach)

We consider the fixed-design regression framework, where one observes

$$Y = (Y_1, \dots, Y_n) = F + \varepsilon \in \mathbb{R}^n \ ,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d., with $\mathbb{E}[\varepsilon_1] = 0$ and $\mathbb{E}[\varepsilon_1^2] = \sigma^2$. The goal is to find from data some $t \in \mathbb{R}^n$ having a small least-squares loss

$$n^{-1} \left\| t - F \right\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - F_i)^2 \ .$$

We then tackle the problem of selecting among several linear estimators, *i.e.*, of the form

$$\widehat{F}_\lambda = A_\lambda Y \ ,$$

where A_λ is a deterministic $n \times n$ matrix. This problem includes:

- model selection for linear regression,
- the choice of a regularization parameter in kernel ridge regression or spline smoothing,
- the choice of a kernel in multiple kernel learning,
- the choice of the number of neighbors (and of a distance in the feature space) for nearest-neighbor regression,
- the choice of a bandwidth (and of a kernel function) for Nadaraya-Watson estimators.

Given a family $(A_\lambda)_{\lambda \in \Lambda}$ of matrices, the goal is to choose some data-driven $\widehat{\lambda} \in \Lambda$ such that the corresponding estimator $\widehat{F}_{\widehat{\lambda}}$ has a quadratic risk $n^{-1}\mathbb{E}\|\widehat{F}_{\widehat{\lambda}} - F\|^2$ as small as possible. When $\mathrm{Card}(\Lambda) \leq K n^\alpha$ for some $K, \alpha \geq 0$, a well-known strategy is to follow the *unbiased risk estimation principle*, *i.e.*, to choose $\widehat{\lambda}$ by minimizing over $\lambda \in \Lambda$ an unbiased estimator of $n^{-1}\mathbb{E}\|\widehat{F}_\lambda - F\|_2^2$. In particular, penalization methods select

$$\widehat{\lambda} \in \arg\min_{\lambda \in \Lambda} \left\{ n^{-1} \|\widehat{F}_\lambda - Y\|_2^2 + \mathrm{pen}(\lambda) \right\} \ , \tag{1}$$

where $\mathrm{pen} : \Lambda \to \mathbb{R}$ is called a penalty. Following the unbiased risk estimation principle, for every $\lambda \in \Lambda$, pen should be close to $n^{-1}\|\widehat{F}_\lambda - F\|_2^2 - n^{-1} \left\| \widehat{F}_\lambda - Y \right\|_2^2$.

Under mild conditions, concentration inequalities show that the risk $n^{-1}\|\widehat{F}_\lambda - F\|_2^2$ and the empirical risk $n^{-1}\|\widehat{F}_\lambda - Y\|_2^2$ both are close to their respective expectation. Therefore, the two key quantities in our problem are

$$\mathbb{E}\left[n^{-1} \|\widehat{F}_\lambda - F\|_2^2 \right] = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} + \frac{\mathrm{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} = \text{bias} + \text{variance} \ , \tag{2}$$

$$\mathbb{E}\left[n^{-1} \left\| \widehat{F}_\lambda - Y \right\|_2^2 \right] = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} - \frac{\left(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n} + \sigma^2 \ . \tag{3}$$

By (2), (3) and the unbiased risk estimation principle, an optimal penalty in (1) would be

$$\text{(4)}\quad \mathrm{pen}_{\mathrm{opt}}(\lambda) = \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] - \mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - Y\right\|_2^2\right] - \sigma^2 = \frac{2\mathrm{tr}(A_\lambda)\sigma^2}{n}\ ,$$

known as Mallows' C_L penalty [7]; its main drawback is its dependence on σ^2, usually unknown. Note that $\mathrm{tr}(A_\lambda)$ is often called *generalized degrees of freedom.*

We extend the notion of *minimal penalty* [4, 3] in order to define an estimator of σ^2 that could be plugged into (4) for designing a fully data-driven penalty. Indeed, let

$$\mathrm{pen}_{\min}(\lambda) = \frac{\left(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n}$$
$$\text{and}\quad \forall C>0,\quad \widehat{\lambda}_{\min}(C) \in \arg\min_{\lambda\in\Lambda}\left\{n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + C\mathrm{pen}_{\min}(\lambda)\right\}\ .$$

By (3), up to concentration inequalities that are detailed in [1, 2], $\widehat{\lambda}_{\min}(C)$ behaves like a minimizer of

$$g_C(\lambda) = \mathbb{E}\left[\frac{\|\widehat{F}_\lambda - Y\|_2^2}{n} + C\mathrm{pen}_{\min}(\lambda)\right] - \sigma^2 = \frac{\|(A_\lambda - I_n)F\|_2^2}{n} + (C-1)\mathrm{pen}_{\min}(\lambda)\ .$$

Therefore, two main cases can be distinguished:

- if $C<1$, then $g_C(\lambda)$ decreases with $\mathrm{tr}(A_\lambda)$ so that $\mathrm{tr}(A_{\widehat{\lambda}_{\min}(C)})$ is huge: $\widehat{\lambda}_{\min}(C)$ overfits.
- if $C>1$, then $g_C(\lambda)$ increases with $\mathrm{tr}(A_\lambda)$ when $\mathrm{tr}(A_\lambda)$ is large enough, so that $\mathrm{tr}(A_{\widehat{\lambda}_{\min}(C)})$ is much smaller than when $C<1$.

As a conclusion, $\mathrm{pen}_{\min}(\lambda)$ is the minimal amount of penalization needed so that a minimizer $\widehat{\lambda}$ of a penalized criterion is not clearly overfitting.

Since $\sigma^{-2}\mathrm{pen}_{\min}(\lambda)$ is known, we deduce the following algorithm:

Input: Λ a finite set with $\mathrm{Card}(\Lambda) \le Kn^\alpha$ for some $K, \alpha \ge 0$, and matrices A_λ.

- $\forall C>0$, compute $\widehat{\lambda}_0(C) = \widehat{\lambda}_{\min}(C\sigma^{-2}) \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + C(2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda))\}$.
- Find $\widehat{C}$ corresponding to the largest jump of $C \to \mathrm{tr}(A_{\widehat{\lambda}_0(C)})$.

Output: $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + 2\widehat{C}\mathrm{tr}(A_\lambda)\}$.

We prove in [1, 2] that if the ε_i are Gaussian, under mild assumptions on the bias term $\|(A_\lambda - I_n)F\|_2^2$, then $|\sigma^{-2}\widehat{C} - 1| \le \kappa\sqrt{\ln(n)}n^{-1/4}$ with large probability, for some constant $\kappa>0$. Furthermore, we deduce that $\widehat{\lambda}$ satisfies an oracle inequality with leading constant $1+\epsilon_n$ on an event of probability at least $1-n^{-2}$.

Previous results on minimal penalties [4, 3, 6] considered the case of projection estimators, for which $\mathrm{tr}(A_\lambda^\top A_\lambda) = \mathrm{tr}(A_\lambda)$, so that the minimal penalty is exactly

half the optimal penalty. Our result shows that for general linear estimators, the optimal and minimal penalties have different shapes, and their ratio

$$\frac{\mathrm{pen}_{\mathrm{opt}}(\lambda)}{\mathrm{pen}_{\mathrm{min}}(\lambda)} = \frac{2\mathrm{tr}(A_\lambda)}{2\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)}$$

can take any value in $(1; 2]$.

Simulation experiments with kernel ridge regression and multiple kernel learning show that the proposed algorithm often improves significantly existing calibration procedures such as 10-fold cross-validation or generalized cross-validation [5], for moderate values of the sample size [1, 2].

References

[1] S. Arlot and F. Bach, *Data-driven calibration of linear estimators with minimal penalties*, In Advances in Neural Information Processing Systems **22** (2009).

[2] S. Arlot and F. Bach, *Data-driven calibration of linear estimators with minimal penalties*, arXiv:0909.1884v1 (2009).

[3] S. Arlot and P. Massart, *Data-driven calibration of penalties for least-squares regression*, Journal of Machine Learning Research **10** (2009), 245–279.

[4] L. Birgé and P. Massart, *Minimal penalties for Gaussian model selection*, Probability Theory and Related Fields, **138** (2007), 33–73.

[5] P. Craven and G. Wahba, *Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numerische Mathematik **31** (1978/79), 377–403.

[6] M. Lerasle, *Optimal model selection in density estimation*, http://hal.archives-ouvertes.fr/hal-00422655/en/ (2009).

[7] C. L. Mallows, *Some comments on C_p*, Technometrics **15** (1973), 661–675.

Estimator selection

Yannick Baraud

Consider a collection $\mathcal{E} = \{\hat{s}_\lambda,\ \lambda \in \Lambda\}$ of arbitrary estimators based on an observation $X \sim P_s$ in view of estimating the parameter s. We propose a selection procedure, based on X as well, which aims at selecting an estimator $\hat{s}_{\hat{\lambda}}$ among $\mathcal{E}$ whose risk is as close as possible to the infimum of those among the collection. The procedure we propose requires little assumption both on s and $\mathcal{E}$, the dependency of the estimators $\hat{s}_\lambda$ with respect to X being possibly unknown. We establish non-asymptotic risk bounds for the selected estimator and show how one can deduce oracle-type inequalities under a posteriori information on the $\hat{s}_\lambda$. The problem of selecting among a given family of estimators arise in many statistical approaches among which model selection, aggregation, construction of robust estimators, etc. The procedure also provides an alternative, at least theoretically, to the resampling techniques such as cross-validation and V-fold, the aim of which is to calibrate a tuning parameter λ. Finally, we show how the procedure can be used in the regression setting for the problem of variable selection, when the number of variables is larger than the number of observations and the errors admit no finite moment.

References

[1] S. Arlot and F. Bach, *Data-driven calibration of linear estimators with minimal penalties*, Technical report, HAL : hal-00414774, version 1 (2009).

[2] A. Barron, , L. Birgé, and P. Massart, *Risk bounds for model selection via penalization*, Probab. Theory Related Fields, **113(3)** (1999), 301 – 413.

[3] L. Birgé, *Model selection via testing: an alternative to (penalized) maximum likelihood estimators*, Ann. Inst. H. Poincaré Probab. Statist., **42(3)** (2006),273–325.

[4] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, *Aggregation and sparsity via l_1 penalized least squares*, In Learning theory, volume **4005** of Lecture Notes in Comput. Sci., Springer, Berlin (2006), 379–391.

[5] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, *Aggregation for Gaussian regression*, Ann. Statist., **35(4)** (2007), 1674–1697.

[6] O. Catoni, *Statistical learning theory and stochastic optimization*, In Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, Springer-Verlag, Berlin (2001).

[7] A. Goldenshluger, *A universal procedure for aggregating estimators*, Ann. Statist., **37(1)** (2009), 542–568.

[8] A. Juditsky, and A. Nemirovski, *Functional aggregation for nonparametric regression*, Ann. Statist., **28(3)** (2000), 681–712.

[9] O. V. Lepski, *A problem of adaptive estimation in Gaussian white noise*, Teor. Veroyatnost. i Primenen., **35(3)** (1990), 459–470.

[10] J.-M. Loubes, and S. van de Geer, *Adaptive estimation with soft thresholding penalties*, Statist. Neerlandica, **56(4)** (2002), 454–479.

[11] P. Rigollet, and A. B. Tsybakov, *Linear and convex aggregation of density estimators*, Math. Methods Statist., **16(3)** (2007), 260–280.

[12] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, **58(1)**(1996), 267–288.

[13] Y. Yang, *Combining different procedures for adaptive regression*, J. Multivariate Anal., **74(1)** (2000), 135–161.

Optimal Rates for Conjugate Gradient Regularization

Gilles Blanchard

(joint work with Nicole Krämer)

Summary We prove optimal rates of convergence (up to a logarithmic factor) in the statistical sense for conjugate gradient (with early stopping) regularization of kernel-based regression problems. (The property of universal consistency of this kind of method was established earlier [1].) The rates are obtained under the assumption that the true regression function belongs to the reproducing kernel Hilbert space. If this assumption is not fulfilled, we obtain convergence rates if additional unlabeled data are available. The rates in these two cases match those obtained by A. Caponnetto [3] for linear regularization operators.

Conjugate Gradient Regularization We observe an i.i.d. sample of n observations $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ where $P(X, Y)$ follows

$$Y_i \quad = \quad f^*(X_i) + \varepsilon_i \, .$$

We assume that the true regression function f^* belongs to the space $\mathcal{L}_2(P_X)$ of square-integrable functions.

We implicitly map the data to a reproducing kernel Hilbert space $\mathcal{H}_k$ with a kernel k. We denote by $K_n = (k(X_i, X_j)) \in \mathbb{R}^{n\times n}$ the kernel matrix and by $\boldsymbol{Y} \in \mathbb{R}^n$ the n centered response observations $Y_1, \ldots, Y_n$.

We propose conjugate gradient (cg) techniques in combination with early stopping for the regularization of the kernel based learning problem. These techniques restrict the learning problem on a nested set of data-dependent subspaces, so-called Krylov subspaces

$$\mathcal{K}_m(\boldsymbol{Y}, K_n) \quad = \quad \text{span}\left\{\boldsymbol{Y}, K_n\boldsymbol{Y}, \ldots, K_n^{m-1}\boldsymbol{Y}\right\} .$$

We define the K_n-norm as $\|\alpha\|_{K_n} = \sqrt{\langle\alpha, K_n\alpha\rangle}$. The cg solution after m iterations is formally defined as

$$\alpha_m \quad = \quad \arg\min_{\alpha\in\mathcal{K}_m(\boldsymbol{Y},K_n)} \|\boldsymbol{Y} - K_n\alpha\|_{K_n}$$

and can be conveniently computed using an iterative formula only involving forward multiplications by K_n. The kernel coefficients α_m define an estimate

$$f_m(X) \quad = \quad \sum_{i=1}^{n} \alpha_{m,i} k(X_i, X)$$

of the true regression function f^*. The number m of cg iterations is the model parameter. We defined in earlier work an early stopping rule for m ensuring universal consistency [1]. The results presented here consider a different stopping rule based on a variation of the discrepancy principle (the latter has been shown by Nemirovski to yield optimal convergence rates for CG in a derministic setting; see [4] for a recent comprehensive account of the topic)

Assumptions The kernel is bounded, $k(x, x') \le \kappa$ for all $x, x' \in \mathcal{X}$, and the noise is bounded, $|\epsilon| \le M$ almost surely. We define the kernel operator

$$K : \mathcal{L}_2(P_X) \to \mathcal{L}_2(P_X),\ g \mapsto \int k(., x')g(x')dP(x') .$$

The regularity of the function f^* is measured in terms of the source condition **SC**(r, ρ): $f^* = K^r u$ with $\|u\| \le \kappa^{-r}\rho$. (In particular, if $r \ge 1/2$, f^* lies in $\mathcal{H}_k$.) The regularity of the kernel operator K is measured in terms of its intrinsic dimensionality **ID**(s, D): There exists $D \ge 1$ such that $\text{Tr}(K(K + \lambda)^{-1}) \le D^2(\kappa^{-1}\lambda)^{-s}$ for all $\lambda \in (0, 1]$.

Case 1: $r \ge 1/2$, which implies $f^* \in \mathcal{H}_k$. We set

$$\lambda_* = \kappa\left(\left(4D/\sqrt{n}\right)\log\left(6/\gamma\right)\right)^{\frac{2}{2r+s}} \text{ for } \gamma > 0$$

and assume n is large enough to ensure $\lambda_* \le \kappa$. For $\tau > 0$, consider the following stopping rule

$$\widehat{m} = \min\left\{m \,\middle|\, \|f_m(X_i) - \boldsymbol{Y}\| \le (2 + \tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\right\} ,$$

where $\delta(\lambda_*) := (3/4)\, M\, (\lambda_*/\kappa)^r$.

Theorem 1. *If conditions* $\mathbf{SC}(r,\rho)$ *and* $\mathbf{ID}(s)$ *are satisfied with* $r \geq 1/2$*, the above stopping rule ensures that, with probability larger than* $1-3\gamma$*,*

$$\|f_{\widehat{m}} - f^*\|_2 \leq c(r,\tau)(M+\rho)\left(\frac{4D}{\sqrt{n}}\log\frac{6}{\gamma}\right)^{\frac{2r}{2r+s}} .$$

Case 2: $r < 1/2$. Similar to the setting studied by Caponnetto, we assume that we have additional unlabeled data of order $n(\lambda^*)^{-(1-2r)}$. After reformulating cg in a semi-supervised setting (using both labeled and unlabeled data), we obtain the same convergence rates as those for linear regularization operators if $r+s \leq 1/2$. The details are omitted.

References

[1] G. Blanchard, N. Krämer, *Kernel Partial Least Squares is Universally Consistent.* JMLR Workshop and Conference Proceedings: AISTATS **9** (2010),57–64.

[2] G. Blanchard, N. Krämer, *Optimal Rates for kernel Conjugate Gradient Regularization.* In preparation.

[3] A. Caponnetto, *Optimal Rates for Regularization Operators in Learning Theory.* CBCL Paper 264/ CSAIL-TR 2006-062, MIT (2006).

[4] M. Hanke, *Conjugate gradient type methods for linear ill-posed problems*, Pitman Research Notes in Mathematics Series, **327** (1995).

Sparse graphs and causal inference

Peter Bühlmann

(joint work with Marloes H. Maathuis and Markus Kalisch)

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

Our work is motivated by the following problem in biology. We want to know which genes play a role for a certain phenotype, say a disease status or the expression of another gene. To be more precise, our goal is to infer which genes have an effect on the phenotype in terms of an intervention: if we knocked down single genes, which of them would show a relevant or important effect on the phenotype? The difficulty is, however, that the available data are only observational.

Using such observational data, we want to infer all (single gene) intervention effects. This task coincides with inferring causal effects, a well-established area in statistics, cf. [1] or [2]. We emphasize that in our applications, it is exactly the intervention or causal effect which is of interest, rather than a regression-type effect of association.

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

Given a set of conditional dependencies from observational data and a corresponding DAG model, one can compute causal effects using intervention calculus [1].

Here, we consider the problem of inferring causal information from observational data, under the assumption that the data were generated by an unknown DAG. This is a more realistic assumption, since in many practical problems, one does not know the DAG. In this scenario, the causal effect is typically not defined uniquely, and that is not surprising given the description of causality by [1] above.

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

We describe a new, computationally feasible algorithm, even if the number of variables (i.e. nodes in the graph) is large, which uses the CPDAG as input for inferring lower bounds on intervention or causal effects. Furthermore, we show that in the case of noise and estimation error, we can still asymptotically infer the CPDAG and the lower bounds for causal effects even if the number of variables p (number of nodes in the graph) is much larger than sample size n, $p \gg n$. Such a consistency result relies on sparsity of the (causal) DAG and the so-called faithfulness assumption for the data-generating probability distribution with respect to the underlying DAG. Details are given in [4] and some of the results there rely on [3]. Furthermore, we validate the method to predict the strongest intervention effects in a large-scale biological system from S.Cerevisiae [5].

References

[1] J. Pearl, *Causality: models, reasoning and inference*, Cambridge University Press (2000).
[2] P. Spirtes and C. Glymour and R. Scheines, *Causation, Prediction, and Search*, 2nd edition, The MIT Press (2000).
[3] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, Journal of Machine Learning Research, **8** (2007), 613 – 636.
[4] M.H. Maathuis and M. Kalisch and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data*, The Annals of Statistics, **37** (2009), 3133 – 3164.
[5] M.H. Maathuis and D. Colombo and M. Kalisch and P. Bühlmann, *Predicting causal effects in large-scale systems from observational data*, Nature Methods, **7** (2010), 247 – 248.

Adaptive nonparametric estimation for several conditional functions

Fabienne Comte

(joint work with Elodie Brunel, Claire Lacour, Stéphane Gaïffas, Agathe Guilloux)

This talk aimed to present results of works by Brunel, Comte and Lacour (2007, 2008), Comte, Gaïffas and Guilloux (2008) and Brunel and Comte (2009). The questions studied in those papers are all related with survival analyis. In such a context indeed, it is natural to introduce covariates in the model, and to look for estimators not only of the hazard rate of the patients, for instance, but also of their hazard rate given their age. This context also leads to wonder if censoring can be taken into account.

Therefore, I explained in my talk how nonparametric bivariate estimators can be built, which are regression estimators in the x-direction (the "covariate-direction") and density (or other type of) estimators in the y-direction. A strategy can be developed to provide definition of collections of nonparametric estimators of the conditional density (or the conditional cumulative distribution function (c.d.f.), the conditional hazard rate, the conditional mean residual life) of Y given $X = x$. These estimators admit developments following product bases, which can be anisotropic. Model selection can then be done in order to keep only a relevant number of coefficients in the decomposition. In all cases, it is performed via contrast penalization. The risk bounds which are obtained for the final estimators are nonasymptotic and follow from Talagrand type inequalities, or Bernstein type inequalities associated with chaining methods. Then, asymptotic anisotropic rates can be deduced and proved to be optimal.

The method can be illustrated through the four examples studied in the aforementioned papers, to show both similarities and differences between them and to derive general principles.
The first idea for the study lies in the way the contrasts are built. For instance in

the conditional density setting we take:

$$\Gamma_n^{(1)}(T) = \frac{1}{n}\sum_{i=1}^{n}\int t^2(X_i,y)dy - \frac{2}{n}\sum_{i=1}^{n} T(X_i,Y_i),$$

and for the conditional c.d.f.

$$\Gamma_n^{(2)}(T) = \frac{1}{n}\sum_{i=1}^{n}\int t^2(X_i,y)dy - \frac{2}{n}\sum_{i=1}^{n}\int T(X_i,y)1_{\{Y_i\leq y\}}dy.$$

It is interesting to study and compare these contrasts, and to understand in what sense the second one is associated with one-dimensional type model selection and rates, in the x-direction only. Note that the definition of minimizers for these contrasts are not always straightforward.

The limit of the strategy is that, when censored variables are considered in this setting, we may obtain results only under the "strong" assumption of independence between the censoring variable and the couple (Y,X) of the variable of interest and the covariate. This is what happens in the case of the conditional density or of the conditional c.d.f. More precisely, assume that the observations are no longer the sequence $(X_i,Y_i)_{1\leq i\leq n}$ but $(X_i,Z_i,\delta_i)_{1\leq i\leq n}$ where $Z_i = Y_i \wedge C_i$ and $\delta_i = 1_{\{Y_i\leq C_i\}}$, which is the standard context of right-censoring. The sequence (C_i) is an i.i.d. sequence, and two types of assumptions are considered concerning the mechanism:

$$\text{(Strong)}\quad (C_i)_{1\leq i\leq n} \text{ independent of } (X_i,Y_i)_{1\leq i\leq n}$$

or

$$\text{(Weak)}\quad (C_i)_{1\leq i\leq n} \text{ independent of } (Y_i)_{1\leq i\leq n} \text{ conditionally to } (X_i)_{1\leq i\leq n}.$$

In the case (Strong), the censoring correction called IPCW (Inverse Probability Censoring Weights) is simple and works well. In the density case for instance, the contrast can simply be modified as follows:

$$\Gamma_n^{(1)}(T) = \frac{1}{n}\sum_{i=1}^{n}\int t^2(X_i,y)dy - \frac{2}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\bar{G}}(Z_i)}T(X_i,Z_i),$$

where $\hat{\bar{G}}(.)$ is the Kaplan Meier estimator of $\bar{G} = 1-G$ the survival function of the C_i's. As this estimator, in the same way as the empirical c.d.f., converges to the true function with paprametric rate, the modification just involves an additional negligible residual term.

Unfortunately, a censoring correction of the same type in the (Weak) case would require to use weights $\delta_i/\hat{\bar{G}}(X_i,Z_i)$ for an estimator $\hat{\bar{G}}(X_i,Z_i)$ of the conditional survival function. Here, the loss is the rate is unavoidable and corresponds to the mean square risk of estimation of $\bar{G}(x,y)$; moreover, in the c.d.f. case, the same thing happens, and it is not relevant to need an estimator of a conditional c.d.f. to estimate another conditional c.d.f.

In fact, to be able to study the (Weak) case, we show, in the case of hazard rate estimation, that the underlying reference measure associated to the problem must

be changed. This is what happens with the contrast

$$\Gamma_n^{(1)}(T) = \frac{1}{n}\sum_{i=1}^{n}\int t^2(X_i,y)1_{\{Z_i\geq y\}}dy - \frac{2}{n}\sum_{i=1}^{n}\delta_i T(X_i,Z_i).$$

Indeed, if $h(x,y)$ denotes the conditional hazard rate, we have:

$$\mathbb{E}(\Gamma_n^{(3)}(T) = \iint_A (T(x,y)-h(x,y))^2 d\mu(x,y) + \iint_A h^2(x,y)d\mu(x,y)$$

where

$$d\mu(x,y) = (1-L(x,y))f_X(x)dxdy$$

with f_X denoting the density of X_1 and $L(x,.)$ the conditional c.d.f. of Z_1 given $X_1 = x$.

Then, an adequate estimator can be found, for direct and anisotropic estimation of the hazard rate, under the weak independence assumption. Nonasymptotic risk bounds can also be proved, as well as optimality asymptotic results.

Consequently, the way to get optimal properties for the estimation of the conditional density $f(x,y)$ starting from $h(x,y)$ under the assumtpion (Weak) is to use the link

$$f(x,y) = h(x,y)\exp(-\int_0^y h(x,u)du)$$

which ensures that h and f have the same regularity. As the link between the functions involves regular functions and the estimation is performed on compact sets, the optimal rate obtained for the estimation of h implies optimal rate for f.

The work on conditional hazard regression is generalized to the study of the estimation of the conditional density of marker-dependent counting processes. In all cases, theoretical estimators are first studied in the case of theoretical penalty functions depending on unknown quantities like the supremum of h on the compact of estimation. In a second time, these unknown terms are replaced by estimators and random penalties are used; in that case, the result are of more asymptotic flavour, to make sure that these random penalties are not too far from their theoretical counterpart.

References

[1] E. Brunel, and F. Comte, *Conditional mean residual life estimation*, Prépublication MAP5 **2009-19** (2009).

[2] E. Brunel, F. Comte, and C. Lacour, *Adaptive estimation of the conditional density in presence of censoring*, Sankhya, **69(4)** (2007), 734–763.

[3] E. Brunel, F. Comte, and C. Lacour, *Minimax estimation of the conditional cumulative distribution function under random censorship*, to appear in Sankhya (2008).

[4] F. Comte, S. Gaïffas, and A. Guilloux, *Adaptive estimation of the conditional intensity of marker-dependent counting processes*, Prépublication MAP5 **2008-16** (2008).

High-dimensionality effects in quadratic programs with linear constraints

Noureddine El Karoui

It is often the case in statistics and various branches of applied mathematics that one wishes to solve optimization problems involving parameters that are estimated from data. It is therefore natural to try to characterize the relationship between the solution of the optimization problem involving estimated parameters (the sample version) and the solution we would get if we knew the actual value of the parameters (the population version). An example of particular interest to some is the classical Markowitz portfolio optimization problem in finance, which is an instance of a quadratic program with linear equality constraints.

I discussed some of these questions in the large dimensional setting when the optimization is performed over vectors of size p, and p is comparable to n, the number of observations we use to get our estimates. From a practical standpoint, this asymptotic setting (p and n go to infinity while p/n does not go to zero) tries to capture the difficulties arising from the fact that we have limited amount of data to estimate the parameters appearing in the problem.

I presented results showing that the high-dimensionality of the data (i.e p/n not small) implies significant and quantifiable risk underestimation, in both the case of quadratic programs with linear equality constraints and linear inequality constraints, when the number of constraints is fixed in the asymptotics. I also considered the question of robustness of the conclusions to various distributional assumptions, focusing on understanding the sensitivity of the results to heavy-tails and time correlation. Finally, I discussed the impact of working with non-independent observations and a significant non-classical failure of the bootstrap. A possible robust correction of the problems was also proposed.

The analysis is based on random matrix theory.

l_1/l_2 penalities

Sara van de Geer

A high-dimensional regression model is one where the number of variables p is much larger than the number of observations n. A popular estimation method is least squares with an l_1 - penalty on the regression coefficients. From a practical point of view however, it is often natural to group variables that "belong together". One may then consider applying an l_1/l_2 penalty proportional to the l_1 - norm of the l_2 - norm of coefficients within groups. We present several versions of this idea, including models with within group structure, and multiple regression models, say panel data models with time dependent (smoothly) related coefficients. We show that l_1/l_2 - penalities lead to sparsity oracle inequalities, assuming compatibility between the l_2 - norm of the regression and the l_1/l_2 - norm of the coefficients.

Selection of kernel density estimators: $\mathbb{L}_p$–risk oracle inequalities

Alexander Goldenshluger
(joint work with Oleg Lepski)

Let X be a random variable in $\mathbb{R}^d$ having density f with respect to the Lebesgue measure. We want to estimate f on the basis of the iid sample $\mathcal{X}_n = (X_1, \dots, X_n)$ drawn from f. By an estimator $\hat{f}$ we mean any measurable function $\hat{f}(t) = \hat{f}(\mathcal{X}_n; t) : (\mathbb{R}^d)^n \times \mathbb{R}^d \to \mathbb{R}$. Accuracy of an estimator $\hat{f}$ is measured by the $\mathbb{L}_s$–risk:

$$\mathcal{R}_s[\hat{f}, f] := \Big[\mathbb{E}\|\hat{f} - f\|_s^q\Big]^{1/q}, \quad s \in [1, \infty), \;\; q \geq 1,$$

where $\mathbb{E}$ is the expectation with respect to the probability measure of the observations $\mathcal{X}_n$. The objective is to develop an estimator with small $\mathbb{L}_s$–risk.

The oracle approach to density estimation is based on selection of estimators or models. Given a family of density estimators $\mathcal{F}$, the goal is to propose a measurable choice, say $\hat{f}$, from the family $\mathcal{F}$ so that for every f from a *large* functional class $\mathbb{F}$ the following *oracle inequality* holds

$$\mathcal{R}_s[\hat{f}; f] \leq C \inf_{\tilde{f} \in \mathcal{F}} \mathcal{R}_s[\tilde{f}; f] + \delta_n, \tag{1}$$

where C is a constant independent of f and n, and the remainder δ_n does not depend on f. Oracle inequalities with "small" remainder term δ_n and constant C close to one are of prime interest; they are key tools for establishing minimax and adaptive minimax results in estimation problems. To the best of our knowledge, inequalities of the type (1) in the context of density estimation were established only in the cases $s = 1$ and $s = 2$ (see, e.g., [2], [4], [1], and [5]).

In this paper we develop a selection procedure for a family of kernel density estimators. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a fixed kernel with $\int K(x)\mathrm{d}x = 1$. Given a bandwidth vector $h = (h_1, \dots, h_d)$ define $V_h := \prod_{i=1}^d h_i$ and consider the kernel estimator of f

$$\hat{f}_h(t) := \frac{1}{nV_h} \sum_{i=1}^n K\left(\frac{t - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(t - X_i). \tag{2}$$

Here by u/v for $u, v \in \mathbb{R}^d$ we mean the coordinate–wise division, and $K_h(\bullet) := V_h^{-1} K(\bullet/h)$. Let $h_{\min} = \big(h_{\min}^{(1)}, \dots, h_{\max}^{(d)}\big)$ and $h_{\max} = \big(h_{\max}^{(1)}, \dots, h_{\max}^{(d)}\big)$ be two vectors in $\mathbb{R}^d$ such that $0 < h_{\min}^{(i)} < h_{\max}^{(i)} \leq 1$, $\forall i$. For brevity, we will write $V_{\min} = V_{h_{\min}}$ and $V_{\max} = V_{h_{\max}}$. Consider the family of kernel estimators

$$\mathcal{F}(\mathcal{H}) := \big\{\hat{f}_h : h \in \mathcal{H}\big\}, \;\; \mathcal{H} := \bigotimes_{i=1}^d \big[h_{\min}^{(i)}, h_{\max}^{(i)}\big].$$

We refer to the problem of selecting an estimator from $\mathcal{F}(\mathcal{H})$ as *the bandwidth selection problem*. This problem is central in the area of density estimation.

We propose a measurable choice $\hat{h} \in \mathcal{H}$ such that the resulting estimator $\hat{f} = f_{\hat{h}}$ satisfies the following oracle inequality

$$\mathcal{R}_s[\hat{f}; f] \;\le\; \big(C(s) + 3\|K\|_1\big) \inf_{h\in\mathcal{H}} \mathcal{R}_s[\hat{f}_h; f] + \delta_{n,s}, \quad \forall f \in \mathbb{F}. \tag{3}$$

If $s \ge 2$ then $\mathbb{F}$ is the set of all probabilty densities uniformly bounded by a constant f_∞, $C(s)$ depends only on K and f_∞, and

$$\delta_n = \kappa_1 (\ln n)^{\kappa_2} n^{1/2} \exp\Big\{ - \frac{\kappa_3}{V_{\max}^{2/s}} \Big\}, \quad V_{\max} := \prod_{i=1}^{d} h_{\max}^{(i)}.$$

for some explicitly given constants κ_i, $i = 1, 2, 3$. It should be emphasized that construction of our selection rule does not require knowledge of f_∞. If $s \in [1, 2)$ then the above oracle inequality holds for *any* density f with $C(s)$ depending on K only and the remainder term

$$\delta_n \;=\; \kappa_1 (\ln n)^{\kappa_2} n^{1/s} \exp\Big\{ - \kappa_3 n^{\frac{2}{s}-1} \Big\}.$$

These results allow to derive adaptive minimax results in a wide variety of density estimation settings. In particular, the selection rule leads to a kernel density estimator that is *adaptive minimax* over a scale of the anisotropic Nikol'ski classes. Minimax estimation of densities from such classes was studied in [3].

Our results are easily extended to more general families of kernel estimators $\mathcal{F}$. Let $\mathcal{K}$ be a class of kernels, and consider the family

$$\mathcal{F}(\mathcal{K}, \mathcal{H}) = \big\{ \hat{f}_{(K,h)} : K \in \mathcal{K}, h \in \mathcal{H} \big\},$$

where $\hat{f}_{(K,h)}$ denotes the estimator given in (2) and associated with kernel $K \in \mathcal{K}$ and bandwidth $h \in \mathcal{H}$. Under metric entropy assumptions on the class $\mathcal{K}$ we show that our selection rule applied to the family $\mathcal{F}(\mathcal{K}, \mathcal{H})$ leads to the estimator $\hat{f}$ that satisfies basically the same oracle inequality (3); now the remainder $\delta_{n,s}$ depends on the entropy of the class $\mathcal{K}$ of kernels.

References

[1] L. Birgé, *Model selection for density estimation with $\mathbb{L}_2$–loss*, `arXiv:0808.1416v2, http://arxiv.org`, (2008).

[2] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation,* (2001). Springer, New York.

[3] I. A. Ibragimov and R. Z. Khas'minskii, *More on estimation of the density of a distribution*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **108** (1981), 72–88 (in Russian).

[4] P. Massart, *Concentration Inequalities and Model Selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Lecture Notes in Mathematics, 1896, (2007). Springer, Berlin.

[5] Ph. Rigollet and A. B. Tsybakov. *Linear and convex aggregation of density estimators*, Math. Methods Statist. **16** (2007), 260–280.

On Universal Oracle Inequalities Related to High Dimensional Linear Models

Yuri Golubev

This talk deals with a classical problem of recovering an unknown vector $\vartheta \in \mathbb{R}^n$ from the noisy data

$$Y = A\vartheta + \sigma\xi,$$

where A is a known $m \times n$ - matrix and $\xi = \big(\xi(1), \dots, \xi(m)\big)^\top$ is a standard white Gaussian noise in $\mathbb{R}^m$ with $\mathbf{E}\xi(k) = 0,\ \mathbf{E}\xi^2(k) = 1,\ k = 1, \dots, m$. The noise level σ is assumed to be known.

We start out by considering the maximum likelihood estimate of ϑ

$$\hat{\vartheta}_0 = \arg\min_{\vartheta \in \mathbb{R}^n} \|Y - A\vartheta\|^2 = (A^\top A)^{-1} A^\top Y,$$

where $\|\bullet\|$ stands for the standard Euclidian norm. Its mean square risk is computed as follows :

$$\mathbf{E}\|\hat{\vartheta}_0 - \vartheta\|^2 = \sigma^2 \sum_{k=1}^{n} \lambda^{-1}(k), \tag{1}$$

where $\lambda(k)$ and $\varphi_k \in \mathbb{R}^n$ are eigenvalues and eigenvectors of $A^\top A$.

In what follows, it is assumed solely that $\lambda(1) \geqslant \lambda(2) \geqslant \cdots \geqslant \lambda(n)$. So, A may be severely ill-posed and Equation (1) reveals the principal difficulty in $\hat{\vartheta}_0$: *Its risk may be very large when n is large or when A has a large condition number.*

The simplest way to improve $\hat{\vartheta}_0$ is to suppress large $\lambda^{-1}(k)$ in (1) with the help of a linear smoother; that is, to estimate ϑ by $H\hat{\vartheta}_0$, where H is a properly chosen $n \times n$ - matrix. In what follows, we deal with smoothing matrices admitting the following representation $H = H_\alpha(A^\top A)$, where $H_\alpha(\lambda)$ is a function $\mathbb{R}^+ \to [0, 1]$ which depends on a regularization parameter $\alpha \in [0, \bar{\alpha}]$ such that

$$\lim_{\alpha \to 0} H_\alpha(\lambda) = 1, \quad \lim_{\lambda \to 0} H_\alpha(\lambda) = 0.$$

So, we estimate ϑ with the help of the following family of linear estimators

$$\hat{\vartheta}_\alpha = H_\alpha(A^\top A)(A^\top A)^{-1} A^\top Y$$

and our main goal is to choose the best estimator within this family, or equivalently, the best regularization parameter α. Note that the mean square risk of $\hat{\vartheta}_\alpha$ is computed as follows :

$$L_\alpha(\vartheta) \stackrel{\text{def}}{=} \mathbf{E}\|\hat{\vartheta}_\alpha - \vartheta\|^2 = \sum_{k=1}^{n} \big[1 - h_\alpha(k)\big]^2 \langle \vartheta, \psi_k \rangle^2 + \sigma^2 \sum_{k=1}^{n} \lambda^{-1}(k) h_\alpha^2(k), \tag{2}$$

where here and below

$$h_\alpha(k) \stackrel{\text{def}}{=} H_\alpha[\lambda(k)], \quad \psi_k \stackrel{\text{def}}{=} A\varphi_k / \|A\varphi_k\| \quad \text{and} \quad \langle \vartheta, \psi_k \rangle \stackrel{\text{def}}{=} \sum_{l=1}^{n} \vartheta(l)\psi_k(l).$$

The heuristical motivation of our approach is based on the idea that a good data-driven regularization should minimize in some sense the risk $L_\alpha(\vartheta)$ (see (2)). To

implement this idea we take the classical way related to the famous principle of unbiased risk estimation which goes back to Akaike (1973). This is why we make use of the empirical risk minimization suggesting to compute data-driven regularization parameters as follows :

$$\hat{\alpha} = \arg\min_{\alpha\in(0,\bar{\alpha}]} R_\alpha[Y, Pen], \tag{3}$$

where

$$R_\alpha[Y, Pen] = \|\hat{\vartheta}_0 - \hat{\vartheta}_\alpha\|^2 + \sigma^2 Pen(\alpha),$$

and $Pen(\alpha) : (0, \bar{\alpha}] \to \mathbb{R}^+$ is a given penalty function. The main difficulty in this approach is related to the choice of the penalty. Intuitively, we want that the method mimics the oracle regularization parameter $\alpha^* = \arg\min_\alpha L_\alpha(\vartheta)$. Therefore we are looking for a minimal penalty ensuring the following inequality

$$L_\alpha(\vartheta) \lesssim R_\alpha[Y, Pen] + \mathcal{C}, \tag{4}$$

where $\mathcal{C} = -\|\vartheta - \hat{\vartheta}_0\|^2$.

A traditional approach to solving (4) is based on the unbiased risk estimation defining the penalty as a root of the equation

$$L_\alpha(\vartheta) = \mathbf{E}R_\alpha[Y, Pen] + \mathbf{E}\mathcal{C}.$$

Is is easily seen that

$$Pen(\alpha) = 2\sum_{k=1}^{n} \lambda^{-1}(k) h_\alpha(k).$$

Unfortunately, in spite of its very natural motivation, this penalty fails for ill-posed inverse problems (see e.g. [2]).

Our main idea is to compute the penalty in a little bit different way, namely, as a minimal function assuring the following inequality

$$\mathbf{E}\sup_{\alpha\leqslant\bar{\alpha}}\Big[L_\alpha(\vartheta) - R_\alpha[Y, Pen] - \mathcal{C}\Big]_+ \leqslant K\mathbf{E}\Big[L_{\bar{\alpha}}(\vartheta) - R_{\bar{\alpha}}[Y, Pen] - \mathcal{C}\Big]_+, \tag{5}$$

where $[x]_+ = \max\{0, x\}$ and $K > 1$ is a constant. The heuristical motivation behind this approach is rather transparent : We are looking for a minimal penalty that balances all excess risks uniformly in $\alpha \in (0, \bar{\alpha}]$.

In the general case, solving (5) is a hard numerical problem, but for the class of ordered smoothers this problem becomes feasible. This class of regularizing matrices $H_\alpha(\bullet)$ is defined, according to Kneip (1994), as follows :

Definition 1. *The family of functions* $\{H_\alpha(\lambda),\ \alpha \in (0, \bar{\alpha}],\ \lambda \in \mathbb{R}^+\}$ *is called ordered smoothers if:*

(1) For any given $\alpha \in (0, \bar{\alpha}]$, $H_\alpha(\lambda) : \mathbb{R}^+ \to [0, 1]$ *is a monotone function.*

(2) If for some $\alpha_1, \alpha_2 \in (0, \bar{\alpha}]$ *and* $\lambda' \in \mathbb{R}^+$,

$$H_{\alpha_1}(\lambda') < H_{\alpha_2}(\lambda')$$

then for all $\lambda \in \mathbb{R}^+$

$$H_{\alpha_1}(\lambda) \leqslant H_{\alpha_2}(\lambda).$$

Note that the class of ordered spectral regularizations is rather vast including *spectral cut-off, Tikhonov-Phillips and Landweber* methods.

Suppose the penalty has the structure

$$Pen(\alpha) = 2\sum_{k=1}^{n} \lambda^{-1}(k) h_\alpha(k) + (1+\gamma)Q(\alpha),$$

where γ is a positive number and $Q(\alpha)$, $\alpha > 0$ is defined as follows :

$$Q(\alpha) = 2D(\alpha)\mu_\alpha \sum_{k=1}^{n} \frac{\rho_\alpha^2(k)}{1 - 2\mu_\alpha \rho_\alpha(k)},$$

where

$$D(\alpha) = \left\{ 2\sum_{k=1}^{n} \lambda^{-2}(k) \big[2h_\alpha(k) - h_\alpha^2(k)\big]^2 \right\}^{1/2}, \quad \rho_\alpha(k) = \frac{\sqrt{2}\big[2h_\alpha(k) - h_\alpha^2(k)\big]}{D(\alpha)\lambda(k)}$$

and μ_α is a root of equation

$$\sum_{k=1}^{n} F[\mu_\alpha \rho_\alpha(k)] = \log \frac{D(\alpha)}{D(\bar\alpha)}, \quad \text{with} \quad F(x) = \frac{1}{2}\log(1-2x) + x + \frac{2x^2}{1-2x}.$$

The following theorem representing the main result in this talk controls the performance of the empirical risk minimization in terms of the penalized oracle risk

$$r(\vartheta) \stackrel{\text{def}}{=} \inf_{\alpha \leqslant \bar\alpha} \bar R_\alpha(\vartheta),$$

where

$$\bar R_\alpha(\vartheta) \stackrel{\text{def}}{=} \mathbf{E}\big\{ R_\alpha[Y, Pen] + \mathcal{C} \big\} = L_\alpha(\vartheta) + (1+\gamma)\sigma^2 Q(\alpha).$$

Theorem 1. *The mean square risk of $\hat\vartheta_{\hat\alpha}$ with $\hat\alpha$ defined by (3) is bounded uniformly in $\vartheta \in \mathbb{R}^n$ by*

$$\mathbf{E}\|\vartheta - \hat\vartheta_{\hat\alpha}\|^2 \leqslant r(\vartheta)\left\{ 1 + \left[\frac{C}{\gamma} \log^{-1/2} \frac{Cr(\vartheta)}{\sigma^2 D(\bar\alpha)} + \frac{C\sigma^2 D(\bar\alpha)}{\gamma^4 r(\vartheta)} \right]^{1/2} \right\},$$

where C is a constant.

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle* Proc. 2nd Intern. Symp. Inf. Theory, Petrov P.N. and Csaki F. eds. Budapest (1973), 267–281.

[2] L. Cavalier and Yu. Golubev, *Risk hull method and regularization by projections of ill-posed inverse problems.* Ann. Statist. **34** (2006), 1653–1677.

[3] A. Kneip, *Ordered linear smoothers*, Ann. Statist. **22** (1994), 835–866.

Higher Criticism thresholding: optimal feature selection when useful features are rare and weak

Jiashun Jin
(joint work with David Donoho)

Consider a two-class classification setting where we have a set of n training samples (Y_i, X_i), $1 \le i \le n$. For each $i = 1, 2, \ldots, n$, Y_i is the class label which equals to 1 if the i-th sample comes from one class and equals to -1 otherwise, $X_i \in R^p$ is the feature vector which is distributed as $N(Y_i \bullet \mu,\ I_p)$ for some unknown contrast mean vector $\mu \in R^p$. Given a new test feature $X \sim N(Y \bullet \mu,\ I_p)$ where the corresponding label Y is unknown, our goal is to predict Y as 1 or -1 with an error as small as possible.

Following our papers [2, 3, 4, 10], we model the coordinates of μ as samples from the mixture of two point masses $(1-\epsilon)\nu_0 + \epsilon\nu_{\mu_0}$. Note that a feature is useful for classification if and only if the corresponding coordinate of μ is nonzero. Let $\tau = \sqrt{n}\mu_0$. Our main interest is in the case where ϵ is small and τ is small or moderately large, so that the useful features are both *rare* and *weak*. We denote such a model by $RW(\epsilon, \tau; n, p)$.

We adopt an asymptotic framework where p tends to ∞ and (ϵ, τ, n) are linked to p through some fixed parameters as p ranges. In detail, fixing a parameter $\beta \in (0, 1)$, we let

$$\epsilon = \epsilon_p = p^{-\beta}.$$

This models the case where the useful features get increasingly rare as p grows. To counter this effect, τ must grow with p. Fixing another parameter $r \in (0, 1)$, we let

$$\tau = \tau_p = \sqrt{2r \log p}.$$

This captures the most interesting range of τ_p: when $r > 1$, the feature selection problem is relatively easy and we can select features by thresholding at $\sqrt{2 \log p}$ (say); when $r \approx 0$, successful classification is impossible.

In addition, we consider three different types of linkage between n and p, where as p tends to ∞, $n = n_p$ may have *no growth*, *slow growth*, or *regular growth*.

- *No Growth.* $n_p = n_0$ for some fixed integer n_0.
- *Slow Growth.* $n_p \to \infty$, but $n_p/p^{\vartheta} \to 0$ for any $\vartheta > 0$.
- *Regular Growth.* $n_p = p^{\vartheta}$ for some $\vartheta \in (0, 1)$.

Combining the above linkages we have the *asymptotic rare/weak model* $ARW(\beta, r, n_p)$. It turns out that, for each of type of growth of n_p, there is an interesting two-phase structure in the region $(\beta, r) \in (0, 1)^2$ which we now describe.

Introduce the *standard phase boundary* function [6, 1, 9]

$$\rho(\beta) = \begin{cases} 0, & 0 < \beta \le 1/2, \\ \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1. \end{cases}$$

For each of the three growth types of n_p, define $\rho^\star(\beta)$ for $\star = N, S, R$ by

$$\rho^N(\beta) = \rho^N(\beta, n_0) = \frac{n_0}{n_0 + 1}\rho(\beta), \qquad 0 < \beta < 1,$$

$$\rho^S(\beta) = \rho(\beta), \qquad 0 < \beta < 1,$$

and

$$\rho^R(\beta) = (1 - \vartheta)\rho(\frac{\beta}{1 - \vartheta}), \qquad 0 < \beta < 1 - \vartheta.$$

In the β-r plane, for each of the three growth types of n_p, we call the region *below* the curve $(\beta, \rho^\star(\beta))$ *Region of Impossibility*, and that *above* the curve *Region of Possibility*.

For each growth type of n_p and a fixed point (β, r) in the interior of Region of Impossibility, consider the sequence of problems $ARW(\beta, r, n_p)$ and a sequence of trained classification methods, perhaps dependent on p. As $p \to \infty$, the misclassification error rate of the resulting trained classifier tends to $1/2$. In this region, the measurement are effectively non-informative, and random guessing does almost as well as any other methods.

At the same time, for each growth type of n_p and a fixed point (β, r) in the interior of Region of Possibility, consider the sequence of problems $ARW(\beta, r, n_p)$. It is possible to have a sequence of trained classification methods, perhaps dependent on p, such that the misclassification error rate tends to 0 as p tends to ∞. In particular, combining Fisher's LDA with a new approach to feature selection (to be described below) yields such a sequence of trained classification methods.

For a p-dimensional weight vector w to be determined, Fisher's LDA classify $\hat{Y} = \mp 1$ according to $L(X) <> 0$, where $L(X)$ is the weighted sum of the test features:

$$L(X) = L(X; w) = \sum_{j=1}^{p} w(j)X(j).$$

Let Z be the summarizing z-vector of the training data:

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i X_i.$$

For a threshold $t > 0$ to be determined, we set weights by

(1)
$$w(j) = w_t^\diamond(j) = \begin{cases} \mathrm{sgn}(Z(j)) \bullet 1_{\{|Z(j)| \geq t\}}, & \diamond = \text{Clipping}, \\ Z(j) \bullet 1_{\{|Z(j)| \geq t\}}, & \diamond = \text{Hard Thresholding}, \\ \mathrm{sgn}(Z(j))(|Z(j)| - t) \bullet 1_{\{|Z(j)| \geq t\}}, & \diamond = \text{Soft Thresholding}. \end{cases}$$

Seemingly, the key for the weight assigning is how to select the threshold t.

In Donoho and Jin (2008, 2009), we select t by Higher Criticism, a notion developed earlier [1] in the context of signal detection. To use Higher Criticism for feature selection, we let $\pi_j = P(|N(0, 1)| \geq |Z(j)|)$ be the p-values associated with

the j-th feature. We then sort the p-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$, and define the j-th Higher Criticism score by

$$HC_{p,j} = \sqrt{p}\left[\frac{j/p - \pi_{(j)}}{\sqrt{j/p(1-j/p)}}\right].$$

Let $\hat{j}^{HC} = \hat{j}^{HC}(Z(1), Z(2), \ldots, Z(p); p)$ be the index at which $HC_{p,j}$ reaches the maximum over all j satisfying $1 \le j \le \alpha_0 p$. We set the threshold for feature selection by the j^{HC}-th largest z-score (in absolute value). We call such a threshold by $t_p^{HC} = t_p^{HC}(Z(1), Z(2), \ldots, Z(p))$. The choice of the threshold is not sensitive to the tuning parameter α_0, which is set as $1/2$ in default.

For each growth type of n_p and a fixed point (β, r) in the interior of Region of Possibility, consider the sequence of problems $ARW(\beta, r, n_p)$. For $\diamond =$ Clipping, Hard Thresholding, or Hard Thresholding, suppose we classify $\hat{Y} = \mp 1$ according to

$$\left(\sum_{j=1}^{p} w_t(j)X(j)\Big|_{\{t=t_p^{HC}\}}\right) \qquad < \quad > 0.$$

Then the misclassification error of the resulting sequence of classification methods tend to 0 as p tends to ∞. The proofs are given in [3] (when n_p has a slow growth) and [4] (all three types of growth of n_p).

References

[1] D. Donoho & J. Jin, *Higher Criticism for detecting sparse heterogeneous mixtures*. Ann. Statist. **32** (2004), 962–994.

[2] D. Donoho & J. Jin, *Higher Criticism thresholding: optimal feature selection when useful features are rare and weak*. Proc. Nat. Acad. Sci. **105(39)** (2008), 14790–14795.

[3] D. Donoho & J. Jin, *Feature selection by Higher Criticism thresholding achieves optimal phase diagram*. Phil. Trans. Roy. Soc. **367** (2009), 4449–4470.

[4] D. Donoho &J. Jin, *When useful features are rare and weak: HCT yields successful classification throughout the region of possibility*. Working manuscript (2010).

[5] P. Hall, Y. Pittelkow & M. Ghosh, *Theoretical measures of relative performance of classifiers for high dimensional data with smal l sample sizes*. J. Roy. Statist. Soc. B **70** (2008), 158–173.

[6] Y.I. Ingster, *Some problems of hypothesis testing leading to infinitely divisible distribution*, Math. Methods Statist. **6** (1997), 47–69.

[7] Y. Ingster, C. Pouet & A.B. Tsybakov, *Classification of sparse high- dimensional vectors*, Phil. Trans. Roy. Soc. **367** (2009a), 4427–4448.

[8] Y. Ingster, C. Pouet & A.B. Tsybakov, *Sparse classification bound- aries*, Eprint arxiv:math/09034807 (2009b).

[9] J. Jin, *Detecting and estimating sparse mixtures*. Ph.D. Thesis, Department of Statistics, Stanford University (2003).

[10] J. Jin, *Impossibility of successful classification when useful features are rare and weak*. Proc. Nat. Acad. Sci. Proc. Nat. Acad. Sci. **106(22)** (2009), 8859–8864.

Sparse Recovery in Infinite Dictionaries

Vladimir Koltchinskii
(joint work with Stas Minsker)

Let (X, Y) be a random couple in $S \times T$, where $(S, \mathcal{A})$ is a measurable space and $T \subseteq \mathbb{R}$ is a Borel set. Let P denote the distribution of (X, Y) and Π denote the distribution of X. Measurable functions $f : S \mapsto \mathbb{R}$ will be called prediction rules. Let $\ell : T \times \mathbb{R} \mapsto \mathbb{R}_+$ be a loss function. Assume that, for all $y \in T$, $\ell(y, \bullet)$ is convex and denote $(\ell \bullet f)(x, y) := \ell(y; f(x))$. The risk of a prediction rule $f : S \mapsto \mathbb{R}$ is defined as

$$P(\ell \bullet f) := \int_{S \times T} (\ell \bullet f) dP = \mathbb{E}\ell(Y; f(X)),$$

the optimal prediction rule is

$$f_* := \operatorname{argmin}_{f:S \mapsto \mathbb{R}} P(\ell \bullet f)$$

and the excess risk of f is

$$\mathcal{E}(f) := P(\ell \bullet f) - P(\ell \bullet f_*).$$

A class $\mathcal{H}$ of measurable functions $h : S \mapsto [-1, 1]$ will be called a dictionary. It will be equipped with a σ-algebra $\mathcal{B}_{\mathcal{H}}$ and with a measure μ. For $\lambda \in L_1(\mu)$, denote

$$f_\lambda(\bullet) := \int_{\mathcal{H}} \lambda(h) h(\bullet) \mu(dh).$$

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sample of i.i.d. copies of (X, Y). Denote P_n the empirical distribution based on this sample (Π_n will denote the empirical distribution based on $(X_1, \dots, X_n)$). We are interested in the following penalized empirical risk minimization problem:

$$\hat{\lambda}^\varepsilon := \operatorname{argmin}_{\lambda \in \mathbb{D}} \left[P_n(\ell \bullet f_\lambda) + \varepsilon \|\lambda\|_{L_1(\mu)} \right], \tag{1}$$

where $\mathbb{D} \subseteq L_1(\mu)$ is a convex set and $\varepsilon > 0$ is a regularization parameter. In the case of finite dictionaries, this is a well known ℓ_1 or LASSO type penalization frequently used in sparse recovery problems. We would like to extend a part of the existing theory for such methods to the case of infinite dictionaries. In what follows, we study only the case of loss functions of quadratic type (including, of course, the loss $\ell(y, u) = (y - u)^2$) and we also assume that $\mathbb{D}$ is a bounded set in $L_1(\mu)$. Our goal is to derive so called sparsity oracle inequalities for the excess risk $\mathcal{E}(f_{\hat{\lambda}^\varepsilon})$ of the solution of (1) in spirit of recent results in the case of finite dictionaries (see, e.g., Bickel, Ritov and Tsybakov [1], Koltchinskii [2, 3], van de Geer [4]).

Alignment coefficients. Let $K : L_2(\mu) \mapsto L_2(\mu)$ denote the following integral operator

$$(Ku)(h) = \int_{\mathcal{H}} \langle h, g \rangle_{L_2(\Pi)} u(g) \mu(dg), h \in \mathcal{H}, u \in L_2(\mu),$$

that can be called the "Gram operator" of the dictionary $\mathcal{H}$. For $w \in L_2(\mu)$, define its alignment coefficient with the dictionary $\mathcal{H}$ as

$$a(w) = a_{\mathcal{H}}(w) = \sup_{\|f_u\|_{L_2(\mu)} \le 1} \langle w, u \rangle_{L_2(\mu)}.$$

It is easy to see that for $w \in \mathrm{Im}(K^{1/2})$, $a(w) = \|K^{-1/2} w\|_{L_2(\mu)}$. In a number of concrete examples, $\mathcal{H} := \{h(t, \bullet) : t \in G\}$, where G is a domain in $\mathbb{R}^d$, and functions on $\mathcal{H}$ can be viewed as functions on G. For smooth functions w, it is often the case that $a(w)$ is dominated by a Sobolev norm of w :

$$a(w) \le C \|w\|_{\mathbb{W}^{2,\alpha}(G)}.$$

This includes, for instance, such dictionaries as $\mathcal{H} = \{\cos\langle t, \bullet \rangle : t \in G\}$, $\mathcal{H} = \{h(\bullet - t) : t \in [0, 2\pi)^d\}$, $\mathcal{H} = \{I_{[0,t]} - I_{(t,1]} : t \in [0,1]\}$ (among others). In such cases, if w is a sum of d "spikes" (i.e., d components that are smooth and have disjoint supports), then $a(w) \le C\sqrt{d}$. Similar bounds on the alignment coefficient hold also when the dictionary $\mathcal{H}$ can be partitioned in a large number N of "almost orthogonal" sets in $L_2(\Pi)$ and w is supported in the union of d of them. In what follows, we denote

$$\partial|\lambda| := \left\{ w : \mathcal{H} \mapsto [-1,1] : w(h) = \mathrm{sign}(\lambda(h)), h \in \mathrm{supp}(\lambda) \right\}, \ \lambda \in \mathbb{D},$$

and we will use the alignment coefficients of functions $w \in \partial|\lambda|$. We will also use the following notation: $S_w := \{h \in \mathcal{H} : |w(h)| \ge 1/2\}, \ w \in \partial|\lambda|$.

Complexity assumptions. We need some complexity assumptions on the dictionary $\mathcal{H}$ that can be expressed in many different ways using random entropy, bracketing entropy, etc. To be specific, let us assume that the following bound on the $L_2(\Pi_n)$-covering numbers holds:

$$\log N(\mathcal{H}; L_2(\Pi_n); u) \le H(u), \ u > 0 \text{ a.s.},$$

where H is a nonnegative nonincreasing function, $H(u) \to \infty, u \to 0$ and H is regularly varying of exponent $\alpha \in [0, 2)$.

Approximation by finite-dimensional subspaces. Given a linear subspace $L \subseteq L_2(\Pi)$ of $\dim(L) < +\infty$ and a subset $\mathcal{H}' \subseteq \mathcal{H}$ of the dictionary, denote

$$\rho(\mathcal{H}'; L) := \sup_{h \in \mathcal{H}'} \|P_{L^\perp} h\|_{L_2(\Pi)},$$

where P_L is the orthogonal projection on a subspace $L \subseteq L_2(\Pi)$ and $L^\perp$ is the orthogonal complement of L.

We will also use the quantity

$$U(L) := \sup_{h \in L, \|h\|_{L_2(\Pi)} \le 1} \|h\|_{L_\infty}$$

that characterizes "smoothness" of functions in L (note that if there exists an $L_2(\Pi)$ orthonormal basis of L of cardinality d, then $U(L) \asymp \sqrt{d}$).

Sparsity Oracle Inequality. We now formulate our main result.

Theorem 1. *There exist constants $C, D > 0$ depending only on ℓ and $\mathbb{D}$ such that the following holds. Let $t > 0$ and denote $t_n := t + 4\log\log_2 n + 2\log 2$. For all $\lambda \in \mathbb{D}$, $w \in \partial|\lambda|$, $L \subseteq L_2(\Pi)$ with $d := \dim(L)$ and $\rho := \rho(S_w; L)$, and for all $\varepsilon \geq D\sqrt{\frac{H(1/\sqrt{d})}{n}}$, with probability at least $1 - e^{-t}$, we have*

$$\mathcal{E}(f_{\hat{\lambda}^\varepsilon}) \leq \mathcal{E}(f_\lambda) + \sqrt{\mathcal{E}(f_\lambda)\xi_n} + \xi_n^2,$$

where

$$\xi_n^2 := C\left[a^2(w)\varepsilon^2 \bigvee \frac{d + t_n}{n} \bigvee \rho\sqrt{\frac{H(\rho/\sqrt{d})}{n}} \bigvee \frac{U(L)H(\rho/\sqrt{d})}{n}\right].$$

Acknowledgment. This research was partially supported by NSF grants MSPA-MCS-0624841, DMS-0906880 and CCF-0808863 at Georgia Institute of Technology.

References

[1] P.J. Bickel, Y. Ritov and A.B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig Selector*, Annals of Statistics **37** (2009), 1705–1732.

[2] V. Koltchinskii, *Sparsity in Penalized Empirical Risk Minimization*, Annales Inst. H. Poincare, Probabilites et Statistique **45** (2009a), 7–57.

[3] V. Koltchinskii, *Sparse Recovery in Convex Hulls via Entropy Penalization*, Annals of Statistics **37** (2009), 1332–1359.

[4] S. van de Geer, *High-dimensional generalized linear models and the Lasso*, Annals of Statistics **36** (2008), 614–645.

Uniform bounds for positive random functionals with application to density estimation

Oleg V. Lepski

The talk consists of three parts. In the first one we presents upper functions for very general stochastic objects namely for positive random functionals. The corresponding results are used for deriving the uniform bounds for gaussian random fields and for the empirical processes. This part is ended by the discussion on the relation of the obtained abstract probabilistic results to the well-known phenomena arising in minimax and minimax adaptive estimation. In the second part we consider some special random processes such that *kernel density estimation process* and *convoluting kernel density estimation process*. Both of them are the special cases of empirical processes. Using the results obtained in the first part of the talk we prove non-asymptotical versions of the law of iterated logarithm and the law of logarithm and compare them with existing asymptotical results. Moreover, we establish also some moments inequalities for the supremum norm of the both mentioned above processes. These results are the crucial tool for the considerations done in the third part of the talk. This part is devoted to statistical

problems and the presented results are subject of the joint work with Alexander Goldenshluger. We study the estimation of a probability density on R^d and consider the risk described by supremum norm. We propose very general selection rule from the family of kernel estimators. The main ingredient of our construction are *majorants* which are the upper functions for the processes considered in the second part. For the selected estimator we prove so-called sup-norm oracle inequality. Being established, an oracle inequality is the informative tool for deriving minimax adaptive results. We use our sup-norm oracle inequality in order to prove that the selected estimator is adaptive over the scale of anisotropic Hölder classes.

Global Uniform Risk Bounds for Wavelet Deconvolution Estimators

Karim Lounici
(joint work with Richard Nickl)

Consider the statistical deconvolution model

$$Y = X + \epsilon$$

where X is a real-valued random variable with unknown probability density $f : \mathbb{R} \to \mathbb{R}^+$ and ϵ is an error term independent of X that is distributed according to the known probability measure φ on $\mathbb{R}$. The law P of Y equals the convolution $f * \varphi$ and we denote its density by g. Let $Y_1, \dots, Y_n$ be i.i.d replications of Y, and denote by P_n the associated empirical measure. The *deconvolution problem* is about recovering the unknown density f from the noisy observations $(Y_1, \dots, Y_n)$. This problem has been extensively studied (see, e.g., [1, 2, 4, 5, 7]).

One key lesson from the above mentioned literature is that some condition on the regularity of the signal ϵ is necessary to be able to estimate f with reasonable accuracy. This condition is often quantified by a lower bound on the decay of the Fourier transform $F[\varphi]$ of φ, and Fourier inversion techniques are applied to construct estimators for f.

Let (φ, ψ) be any scaling and wavelet functions such that $\varphi, \psi \in L^p(\mathbb{R})$ for every $1 \leqslant p \leqslant \infty$, and for some $0 < a' < a$ we have $\mathrm{supp}(F[\varphi]) \subseteq [-a, a]$ as well as $\mathrm{supp}(F[\psi]) \subseteq [-a, -a] \setminus [-a', a']$. We shall assume furthermore that $\sup_{x\in\mathbb{R}} \sum_k |\varphi(x-k)| < \infty$, $\sup_{x\in\mathbb{R}} \sum_k |\psi(x-k)| < \infty$. These conditions are satisfied for Meyer wavelets. Assume the density f admits the formal decomposition $f = \sum_{k\in\mathbb{Z}} \alpha_{jk}\varphi_{jk} + \sum_{l=j}^{\infty} \sum_{k\in\mathbb{Z}} \beta_{lk}\psi_{lk}$, $\alpha_{jk} = \langle f, \varphi_{jk}\rangle$, $\beta_{lk} = \langle f, \psi_{lk}\rangle$ where $\forall u, v \in L^2(\mathbb{R})$ $\langle u, v\rangle = \int_{\mathbb{R}} u(x)\overline{v(x)}dx$.

For any integer $j \geqslant 0$ consider the estimator $f_n(x, j) = \frac{1}{n}\sum_{m=1}^{n} K_j^*(x, Y_m)$ where

$$K_j^*(x, y) = 2^j \sum_{k\in\mathbb{Z}} \varphi(2^j x - k)\tilde{\varphi}_{jk}(y), \quad \tilde{\varphi}_{jk}(y) = F^{-1}\left[2^{-l}\frac{F[\varphi_{0k}](2^{-l}\bullet)}{\overline{F[\varphi]}}\right](y).$$

Note that the above estimator is well-defined if we assume $|F[\varphi](t)| > 0$ on $[-2^j a, 2^j a]$. Define

$$\delta_j = \min_{t\in[-2^j a, 2^j a]} |F[\varphi](t)|.$$

Set $j' = j \vee 1$. Assume that f is bounded. Then there exists a constant $c = c(\varphi, \psi) > 0$ such that $\forall n \geqslant 1$

$$\mathbb{E}\sup_{x\in\mathbb{R}} |f_n(x,j) - Ef_n(x,j)| \leqslant \frac{c}{\delta_j}\left(\|g\|_\infty^{1/2}\sqrt{\frac{2^j j'}{n}} + \frac{2^j j'}{n}\right).$$

If $n \geqslant c' 2^j j'$ for some $c' > 0$ then we have $\forall j \geqslant 0$ and $u > 0$

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}} |f_n(x,j) - Ef_n(x,j)| > \frac{C}{\delta_j}\left(G\sqrt{(1+u)\frac{2^j j'}{n}} + (1+u)\frac{2^j j'}{n}\right)\right\} \leqslant e^{-(1+u)j'}$$

where $G = \max(\|g\|_\infty^{1/2}, 1)$ and $C = C(\varphi, \psi, c') > 0$.

The originality of our approach appears in the computation of an upper bound on the entropy of the class $\{\delta_j \tilde{\varphi}_{jk},\, k \in \mathbb{R}\}$. We combine recent results on VC-property of functions of quadratic variation [3] with Paley-Littlewood theory and the fact that wavelet bases are compatible with both the L^2 and L^∞ structure simultaneously. Note that our results also cover the standard density estimation problem (in this case $\varphi \equiv 0$ and consequently $\delta_j \equiv 1$) and improve upon [3] where some moment condition was imposed on the density.

The above uniform deviation results can be readily applied to derive the optimal rates of convergence for densities $f \in B^s_{\infty,\infty}(\mathbb{R})$.

The Besov ball $B^s_{\infty\infty}(L)$ is defined as the set of functions f such that $f \in L^\infty(\mathbb{R})$ and

$$\|f\|_{s,\infty,\infty} = \|\alpha_{0\bullet}\|_\infty + \max_{l\geqslant 0}\left\{2^{l(s+1/2)}\|\beta_{l(\bullet)}\|_\infty\right\} \leqslant L.$$

Assume $|F[\varphi](t)| \geqslant C(1+|t|^2)^{-\frac{w}{2}} e^{-c_0|t|^\alpha}$, $\forall t \in \mathbb{R}$ where $C, \alpha > 0$ and $c_0, w \geqslant 0$. Consider the linear estimator $f_n(\bullet, j_n)$ with resolution j_n taken such that

$$2^{j_n} \asymp \begin{cases} \left(\frac{n}{\log n}\right)^{\frac{1}{2(s+w)+1}} & \text{if } c_0 = 0 \\ (\tau \log_2 n)^{\frac{1}{\alpha}}, & \text{with } c_0 a^\alpha \tau < 1/2 \text{ if } c_0 > 0. \end{cases}$$

Then there exists a constant $C' = C'(s, L, \varphi, \psi, C, w, c_0, \alpha) > 0$ such that $\forall n \geqslant 2$

$$\sup_{f\in B(s,L)} \mathbb{E}\sup_{x\in\mathbb{R}} |f_n(x, j_n) - f(x)| \leq C' \begin{cases} \left(\frac{1}{\log n}\right)^{\frac{s}{\alpha}} & \text{if } c_0 > 0 \\ \left(\frac{\log n}{n}\right)^{\frac{s}{2s+2w+1}} & \text{if } c_0 = 0 \end{cases}$$

We established our result under the minimal condition on $|F[\varphi](t)|$ to be bounded from below on growing intervals $[-2^j a, 2^j a]$ and no condition whatsoever on the support of f. Note also that the rate we derived are the minimax rates of sup-norm deconvolution.

In the moderately ill-posed case $c_0 = 0$, the optimal choice of the resolution j_n depends on the unknown regularity s of f. We show that the thresholded estimator defined below is minimax adaptive on the Besov balls $B^s_{\infty\infty}(L)$:

$$f_n^T(y) = f_n(y,0) + \sum_{l=0}^{j_1-1} \sum_k \hat{\beta}_{lk} 1_{|\hat{\beta}_{lk}|>\tau} \psi_{lk}(y),$$

where

$$\hat{\beta}_{lk} = \frac{2^{l/2}}{n} \sum_{m=1}^{n} \tilde{\psi}_{lk}(Y_m), \quad \tilde{\psi}_{lk}(y) = F^{-1}\left[2^{-l} \frac{F[\psi_{0k}(2^{-l}\bullet)]}{\overline{F[\varphi]}} \right](y),$$

$\tau = \kappa 2^{wl} G \sqrt{\frac{\log n}{n}}$, $2^{j_1} \asymp \left(\frac{n}{\log n}\right)^{1/(2w+1)}$, $j_1 > 0$ and $\kappa > 0$ is a numerical constant sufficiently large.

Assume that $|F[\varphi](t)| \geqslant C(1+|t|^2)^{-\frac{w}{2}}$ for all $t \in \mathbb{R}$ where $C > 0$,$w \geqslant 0$. Then we have for every $n \geqslant 2$ and every $s > 0$

$$\sup_{f \in B^s_{\infty\infty}(L)} \mathbb{E} \sup_{y \in \mathbb{R}} |f_n^T(y) - f(y)| \leq D \left(\frac{\log n}{n} \right)^{\frac{s}{2(w+s)+1}}$$

where $D > 0$ depends only on $L, \varphi, \varphi, \psi$. This result was established without any moment condition on f and covers the case of standard density estimation.

Note finally that our results can be applied to derive adaptive confidence bands.

References

[1] N. Bissantz, L. Dumbgen, H. Holzmann and A. Munk, *Non-parametric confidence bands in deconvolution density estimation*, J. R. Stat. Soc. Ser. B Stat. Methodol. **69** (2007), 483–506.

[2] C. Butucea and A.B. Tsybakov, *Sharp optimality in density deconvolution with dominating bias. I*, Theory Probab. Appl. **52** (2008), 24–39.

[3] E. Gine and R. Nickl, *Uniform limit theorems for wavelet density estimators*, Ann. Probab. **37** (2009), 1605–1646.

[4] I.M. Johnstone, G. Kerkyacharian, D. Picard and M. Raimondo, *Wavelet deconvolution in a periodic setting*, J. R. Stat. Soc. Ser. B Stat. Methodol. **66** (2004), 547–573.

[5] I.M. Johnstone and M. Raimondo, *Periodic boxcar deconvolution and Diophantine approximation*, Ann. Statist. **32** (2004), 1781–1804.

[6] K. Lounici and R. Nickl, *Global Uniform Risk Bounds for Wavelet Deconvolution Estimators*, Submitted.

[7] M. Pensky and B. Vidakovic, *Adaptive wavelet estimator for nonparametric density deconvolution*, Ann. Statist. **27** (1999), 2033–2053.

Nonparametric Regression on a Generated Covariate with an Application to Semiparametric GARCH-in-Mean Models

Enno Mammen
(joint work with Christian Conrad)

We consider time series models in which the conditional mean of a response variable Y_t given the past $\mathcal{F}_t$ depends on an unobserved covariate h_t. More precisely, we assume that for $t = 1, ..., T$:

$$Y_t = m_0(h_t) + \varepsilon_t,$$

where ε_t fulfills $\mathbf{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0$ for an increasing σ-field $\mathcal{F}_t$ with the property that (ε_t, h_{t+1}) is $\mathcal{F}_t$-measurable.

The covariate process h_t is an unobserved one-dimensional process. We assume that h_t can be consistently estimated by known functions $\widehat{h}_t$ that depend on parameters ψ and m and on the past observations $Y_{t-1}, Y_{t-2}, \ldots$. We denote the true parameter values by ψ_0 and m_0, i.e. $h_t = \widehat{h}_t(\psi_0, m_0)$. A typical example could be that h_t follows a GARCH$(1,1)$ process or another specification from the GARCH family. Most importantly, we allow h_t to depend on the function m_0. In particular, this is the case if $\widehat{h}_t$ depends on ψ_0 and on the residuals $\varepsilon_1, ..., \varepsilon_{t-1}$. Then one needs the function m_0 to get the residuals from the observations $Y_{t-1}, Y_{t-2}, \ldots$. Our central assumption on $\widehat{h}_t$ is that this function is measurable with respect to $\mathcal{F}_{t-1}$.

We discuss estimation of m and testing parametric specifications of m.

Our testing procedure is based on iterative fits of the covariate and nonparametric kernel smoothing of the conditional mean function. The test statistic is given by the L_2 norm of the difference between the kernel estimator and the parametric estimator of m. We show that this test statistic is asymptotically normal and discuss its asymptotic power.

For estimation we consider nonparametric quasi-maximum likelihood sieve estimators of m and ψ:

$$(\widehat{m}, \widehat{\psi}) = \arg \min_{(m,\psi)\in\mathfrak{M}_n} \sum_{t=1}^{T} \ln(h_t(m,\psi)) + \frac{(Y_t - m(h_t(m,\psi)))^2}{h_t(m,\psi)},$$

where $\mathfrak{M}_n$ is an increasing class (sieve) of parameters. We assume entropy conditions on the class $\mathfrak{M}_n$. Under these assumptions we show the following result: $(\widehat{m}, \widehat{\psi})$ converges to (m, ψ) with rate $o_P(n^{-\alpha-\delta})$ for all $\delta > 0$, where α is the optimal rate in a nonparametric regression problem with regression class that has the same entropy as $\mathfrak{M}_n$. This shows that the function m_0 can be estimated with nearly the same rate as if the unobserved covariate h_t would be known.

The proofs of our results are based on empirical process methods.

We apply our approach for testing economic theories that postulate functional relations between macroeconomic or financial variables and their conditional second moments. We illustrate the usefulness of the methodology by testing the linear risk-return relation predicted by the ICAPM.

In a related paper Mammen, Rothe and Schienle (2010) consider a nonparametric regression model:

$$Y = m_0(R) + \varepsilon,$$
$$E[\varepsilon|R] = 0$$

with one-dimensional response Y and q-dimensional covariate R where again the covariate R is unobserved but a nonparametric estimator $\widehat{R}$ of R is available. In that paper the question is studied how the outcome of an estimator of m_0 changes if one regresses on $\widehat{R}$ instead of regressing on R. Leading examples in that paper are control variable approaches for nonparametric regression models with endogenous covariates.

References

[1] C. Conrad, E. Mammen, *Nonparametric Regression on a Generated Covariate with an Application to Semiparametric GARCH-in-Mean Models*, preprint, university of Mannheim (2010).
[2] E. Mammen, C. Rothe, M. Schienle, *Nonparametric Regression with Nonparametrically Generated Covariates*, preprint, university of Mannheim (2010).

Finite-Sample Confidence Bands in Density Estimation

Richard Nickl

Let $X_1, ..., X_n$ be a random sample from some unknown probability density f defined on the unit sphere $\mathbb{S}^d$ of $\mathbb{R}^{d+1}, d \geqslant 1$. Consider the needlet frame $\{\varphi_{j\eta}\}$ describing the needlet projection onto the space of spherical polynomials of degree less than 2^j. We prove non-asymptotic concentration inequalities for the uniform deviations of the linear needlet density estimator $f_n(j)$ obtained from an empirical estimate of the needlet projection $\sum_\eta \varphi_{j\eta} \int f\varphi_{j\eta}$ of f. We apply these results to construct nonasymptotic confidence bands for the unknown density f. The confidence bands are shown to be adaptive over classes of differentiable and Hölder-continuous functions on $\mathbb{S}^d$ that attain their Hölder exponents on $\mathbb{S}^d$. As a byproduct of independent interest we obtain a characterization of Hölder function spaces on $\mathbb{S}^d$ by the needlet approximation spaces.

References

[1] E. Giné, R. Nickl, *Confidence Bands in Density Estimation*, Ann. Statist. **38** (2010), 1122–1170.
[2] G. Kerkyacharian, R. Nickl, D. Picard, *Concentration Inequalities and Confidence Bands for Needlet Density Estimators on the Unit Sphere*, preprint (2010).

Asymptotic equivalence for inference on the quadratic variation of Gaussian martingales

Markus Reiss

In recent years volatility estimation from high-frequency data has attracted a lot of attention in financial econometrics and statistics. Due to empirical evidence that the observed transaction prices of assets cannot follow a semi-martingale model, a prominent approach is to model the observations as the superposition of the true (or efficient) price process with some measurement error, conceived as microstructure noise. The main features are already present in the basic model of observing

$$Y_i = X_{i/n} + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

with an efficient price process $X_t = \int_0^t \sigma(s)\, dB_s$, B a standard Brownian motion, and $\epsilon_i \sim N(0, \delta^2)$ all independent. The aim is to perform statistical inference on the volatility function $\sigma : [0,1] \to \mathbb{R}^+$, e.g. estimating the so-called integrated volatility $\int_0^1 \sigma^2(t)\, dt$ over the trading day.

The mathematical foundation on the parametric formulation of this model has been laid by [2] who prove the interesting result that the model is locally asymptotically normal (LAN) as $n \to \infty$, but with the unusual rate $n^{-1/4}$, while without microstructure noise the rate is $n^{-1/2}$. Starting with [11], the nonparametric model has come into the focus of research. Mainly three different, but closely related approaches have been proposed afterwards to estimate the integrated volatility: multi-scale estimators [10], realized kernels or autocovariances [8] and preaveraging [4]. Under various degrees of generality, especially also for stochastic volatility, all authors provide central limit theorems with convergence rate $n^{-1/4}$ and an asymptotic variance involving the so-called quarticity $\int_0^1 \sigma^4(t)\, dt$. Recently, also the problem of estimating the spot volatility $\sigma^2(t)$ itself has found some interest [6].

The aim of the present work is to provide a thorough mathematical understanding of the basic model, to explain why statistical inference is not so canonical and to propose a simple estimator of the integrated volatility which is efficient. To this end we employ Le Cam's concept of asymptotic equivalence between experiments. In fact, our main theoretical result states under some regularity conditions that observing (Y_i) in (1) is for $n \to \infty$ asymptotically equivalent to observing the Gaussian shift experiment

$$dY_t = \sqrt{2\sigma(t)}\, dt + \delta^{1/2} n^{-1/4}\, dW_t, \quad t \in [0,1],$$

with Gaussian white noise dW. Not only the large noise level $\delta^{1/2} n^{-1/4}$ is apparent, but also a non-linear $\sqrt{\sigma(t)}$-form of the signal, from which optimal asymptotic variance results can be derived. Note that a similar form of a Gaussian shift was found to be asymptotically equivalent to nonparametric density estimation [7]. A key ingredient of our asymptotic equivalence proof are the results by [3] on asymptotic equivalence for generalized nonparametric regression, but also ideas

from [1] and [9] play a role. Moreover, fine bounds on Hellinger distances for Gaussian measures with different covariance operators turn out to be essential.

Roughly speaking, asymptotic equivalence means that any statistical inference procedure can be transferred from one experiment to the other such that the asymptotic risk remains the same, at least for bounded loss functions. Technically, two sequences of experiments $\mathscr{E}^n$ and $\mathscr{G}^n$, defined on possibly different sample spaces, but with the same parameter set, are asymptotically equivalent if the Le Cam distance $\Delta(\mathscr{E}^n, \mathscr{G}^n)$ tends to zero. For $\mathscr{E}_i = (\mathscr{X}_i, \mathscr{F}_i, (\mathbb{P}^i_\vartheta)_{\vartheta\in\Theta})$, $i = 1, 2$, by definition, $\Delta(\mathscr{E}_1, \mathscr{E}_2) = \max(\delta(\mathscr{E}_1, \mathscr{E}_2), \delta(\mathscr{E}_1, \mathscr{E}_2))$ holds in terms of the deficiency $\delta(\mathscr{E}_1, \mathscr{E}_2) = \inf_M \sup_{\vartheta\in\Theta} \left\|MP^1_\vartheta - P^2_\vartheta\right\|_{TV}$, where the infimum is taken over all randomisations or Markov kernels M from $(\mathscr{X}_1, \mathscr{F}_1)$ to $(\mathscr{X}_2, \mathscr{F}_2)$, see e.g. [5] for details. In particular, $\delta(\mathscr{E}_1, \mathscr{E}_2) = 0$ means that $\mathscr{E}_1$ is more informative than $\mathscr{E}_2$ in the sense that any observation in $\mathscr{E}_2$ can be obtained from $\mathscr{E}_1$, possibly using additional randomisations. Here, we shall always explicitly construct the transformations and randomisations and we shall then only use that $\Delta(\mathscr{E}_1, \mathscr{E}_2) \leqslant \sup_{\vartheta\in\Theta} \left\|P^1_\vartheta - P^2_\vartheta\right\|_{TV}$ holds when both experiments are defined on the same sample space.

The asymptotic equivalence is deduced stepwise. The regression-type model (1) is shown to be asymptotically equivalent to a corresponding white noise model with signal X. Then a very simple construction yields a Gaussian shift model with signal $\log(\sigma^2(\bullet) + c)$, $c > 0$ some constant, which is asymptotically less informative, but only by a constant factor in the Fisher information. Inspired by this construction, we present a generalization where the information loss can be made arbitrarily small (but not zero), before applying nonparametric local asymptotic theory to derive asymptotic equivalence with our final Gaussian shift model for shrinking local neighborhoods of the parameters. The global result is based on an asymptotic sufficiency result for simple independent statistics.

References

[1] Carter, Andrew, *A continuous Gaussian process approximation to a nonparametric regression in two dimensions*, Bernoulli, **12(1)** (2006), 143–156.

[2] Gloter, Arnaud and Jacod, Jean, *Diffusions with measurement errors. I: Local asymptotic normality,* ESAIM, Probab. Stat., **5** (2001), 225–242.

[3] Grama, Ion and Nussbaum, Michael, *Asymptotic equivalence for nonparametric regression*, Math. Methods Stat. **11(1)** (2002), 1 – 36.

[4] Jacod, Jean and Li, Yingying and Mykland, Per A. and Podolskij, Mark and Vetter, Mathias, *Microstructure noise in the continuous case: the pre-averaging approach.* Stochastic Processes Appl, **119(7)** (2009), 2249–2276.

[5] Le Cam, Lucien and Yang, Grace Lo, *Asymptotics in statistics. Some basic concepts. 2nd ed.*, Springer Series in Statistics (2000). New York, Springer.

[6] Munk, Axel and Schmidt-Hieber, Johannes, *Nonparametric estimation of the volatility function in a high-frequency model corrupted by noise*, arXiv:0908.3163v2 (2009).

[7] Nussbaum, Michael, *Asymptotic equivalence of density estimation and Gaussian white noise*, Ann. Stat., **24(6)** (1996), 2399–2430.

[8] Ole E. Barndorff-Nielsen and Peter Reinhard Hansen and Asger Lunde and Neil Shephard, *Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise,* Econometrica, **76(6)** (2008), 1481–1536.

[9] Reiß, Markus, *Asymptotic equivalence for nonparametric regression with multivariate and random design*, Ann. Stat., **36(4)** (2008), 1957–1982.
[10] Zhang, Lan, *Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach*, Bernoulli **12(6)** (2006), 1019 – 1043.
[11] Zhang, Lan and Mykland, Per A. and Aït-Sahalia, Yacin, *A tale of two time scales: Determining integrated volatility with noisy high-frequency data* J. Am. Stat. Assoc., **100(472)** (2005), 1394 – 1411.

Adaptive test of homogeneity for Poisson processes when the alternative belongs to Weak Besov bodies

Patricia Reynaud-Bouret
(joint work with Magalie Fromont, Béatrice Laurent)

The presentation consists in a comparison between adaptive test and adaptive estimation for Poisson processes. We observe a Poisson process N with unknown intensity $s(x)$ wrt Ldx on $[0,1]$ where L is a known constant to derive rates of convergence. We assume that $\|s\|_\infty < \infty$ and that one can decompose s on the Haar basis :

$$s = \alpha_0\varphi_0 + \sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1}\alpha_{(j,k)}\varphi_{(j,k)},$$

with $\varphi_0(x) = \mathbf{1}_{[0,1]}(x)$ and $\varphi_{(j,k)}(x) = 2^{j/2}\psi(2^jx-k)$ where $\psi(x) = \mathbf{1}_{[0,1/2[}(x) - \mathbf{1}_{[1/2,1[}(x)$. We want either to estimate s or to test H_0: "s is constant" (ie N is homogeneous) against H_1: "s is not constant".
First let us understand what happens on one finite vectorial subspace. Let $\Lambda \subseteq \{(j,k), j \geq 0, k = 0, ..., 2^j - 1\}$ and $S_\Lambda = \mathrm{Span}(\varphi_0, \varphi_\lambda, \lambda \in \Lambda)$. The dimension of S_Λ is denoted D_Λ. The least-square estimator is defined by

$$\hat{s}_\Lambda = \widehat{\alpha}_0\varphi_0 + \sum_{\lambda\in\Lambda}\widehat{\alpha}_\lambda\varphi_\lambda,$$

with $\widehat{\alpha_\lambda} = \frac{1}{L}\int_{[0,1]}\varphi_\lambda(x)dN_x$. Let s_Λ the orthogonal projection of s on S_Λ, then the risk if this estimator satisfies

$$\mathbb{E}(\|s-\hat{s}_\Lambda\|^2) \leq \|s-s_\Lambda\|^2 + \frac{D_\Lambda\|s\|_\infty}{L}.$$

When we want to test the homogeneity, we actually want to reject when the distance between s and $S_0 = \mathrm{Span}(\varphi_0)$ is too large. The procedure is consequently decomposed as follows:

(1) We approximate $d(s,S_0)^2$ by $\sum_{\lambda\in\Lambda}\alpha_\lambda^2$.
(2) We estimate it unbiasly by $T_\Lambda = \sum_{\lambda\in\Lambda}T_\lambda$ with

$$T_\lambda = \widehat{\alpha}_\lambda^2 - \frac{1}{L^2}\int\varphi_\lambda^2 dN.$$

(3) Under H_0 the law of T_Λ given that $N_{[0,1]} = n$ is free of s, so there exists $t^{(n)}_{\Lambda,\alpha}$ such that

$$\mathbb{P}(T_\Lambda > t^{(n)}_{\Lambda,\alpha} | N_{[0,1]} = n) \leq \alpha.$$

(4) We consequently reject when $T_\Lambda > t^{(N_{[0,1]})}_{\Lambda,\alpha} = t^{(N)}_{\Lambda,\alpha}$.

(5) One possible choice is $t^{(n)}_{\Lambda,\alpha} = q^{(n)}_{\Lambda,\alpha}$ the $1-\alpha$ quantile of the conditional distribution.

The performance of the test is measured in term of separation distance, i.e. the question is: under H_1, how far from S_0 should s be to obtain $\mathbb{P}(\text{accept } H_0) \leq \beta$? If $\mathbb{P}(t^{(N)}_{\Lambda,\alpha} \geq A_{\Lambda,\alpha,\beta}) \leq \beta/3$, and if

$$d^2(s, S_0) \geq \|s - s_\Lambda\|^2 + \square_{\beta,\|s\|_\infty} \frac{\sqrt{D_\Lambda}}{L} + A_{\Lambda,\alpha,\beta},$$

then the error of second kind is less than β. Remark that with respect to the estimation part $\sqrt{D_\Lambda}$ is replacing D_Λ. Test are consequently usually thought to be easier than the corresponding estimation. However the presence of $A_{\Lambda,\alpha,\beta}$ is crucial.
If $t^{(N)}_{\Lambda,\alpha} = q^{(N)}_{\Lambda,\alpha}$,

$$A_{\Lambda,\alpha,\beta} = \square_{\beta,\|s\|_\infty} \left[\frac{\sqrt{D_\Lambda \log(\alpha^{-1})}}{L} + \frac{\log(\alpha^{-1})}{L} + \frac{E_\Lambda \log^2(\alpha^{-1})}{L^2} \right],$$

where $E_\Lambda = \sum_{j/(j,k)\in\Lambda} 2^j$, which may be much larger than D_Λ.
The next step consists in combining tests or estimators. Whereas the estimation in a collection of S_Λ may be difficult (but classical), the testing procedure in a family of tests is quite easy. Let $\mathcal{M}$ be a collection of possible Λ's. Then one rejects H_0 when there exists one $\Lambda \in \mathcal{M}$ such that $T_\Lambda > t^{(N)}_{\Lambda,\alpha_\Lambda} = q^{(N)}_{\Lambda,\alpha_\Lambda}$, where under H_0, $\mathbb{P}(\exists \Lambda \in \mathcal{M},\ T_\Lambda > t^{(N)}_{\Lambda,\alpha_\Lambda}) \leq \alpha$.
The basic choice for α_Λ is the Bonferroni choice ie $\alpha_\Lambda = \alpha/|\mathcal{M}|$. This allow us to define a "nested collection of tests" ie $\mathcal{M} = \{\Lambda_1, ..., \Lambda_{\bar{J}}\}$, with $\Lambda_J = \{(j,k), j \leq J, k = 0, ..., 2^j - 1\}$ whose separation distance is at least

$$\inf_J \|s - s_{\Lambda_J}\|^2 + \square_{\alpha,\beta,\|s\|_\infty} \left[\frac{\sqrt{2^J \bar{J}}}{L} + \frac{\bar{J}}{L} + \frac{2^J \bar{J}^2}{L^2} \right].$$

In the same way, the thresholding estimation procedure has a "test" version which is defined by the following rule: one rejects H_0 when there exists $\lambda \in \Lambda_{\bar{J}}$ such that $(\lambda = (j,k))$

$$T_\lambda > q^{(N)}_{\lambda,\alpha/(2^j \bar{J})}.$$

This is equivalent to ask if there exists $\Lambda \subseteq \Lambda_{\bar{J}}$ such that

$$\sum_{\lambda\in\Lambda} T_\lambda = T_\Lambda > \sum_{\lambda\in\Lambda} q^{(N)}_{\lambda,\alpha/(2^j \bar{J})} = t^{(N)}_{\Lambda,\alpha}.$$

Then the separation distance of this test is at least

$$\inf_{\Lambda\subseteq\Lambda_{\bar J}}\|s-s_\Lambda\|^2+\square_{\alpha,\beta,\|s\|_\infty}\left[\frac{D_\Lambda\bar J}{L}+\frac{D_\Lambda\bar J^2 2^{\bar J}}{L^2}\right].$$

One can see the difference between both tests by looking at the minimax separation rate over $\mathcal{B}^\delta_{2,\infty}(R)\cap W_\gamma(R')$ where the classical Besov body is defined by

$$\mathcal{B}^\delta_{2,\infty}(R)=\left\{s\geq 0\ \Big/\ \forall j\in\mathbb{N},\sum_{k=0}^{2^j-1}\alpha^2_{(j,k)}\leq R^2 2^{-2j\delta}\right\}$$

and the weak Besov body is defined by

$$W_\gamma(R')=\left\{s\geq 0\ \Big/\ \forall t>0,\sum_{j\in\mathbb{N}}\sum_{k=0}^{2^j-1}\alpha^2_{(j,k)}\mathbf{1}_{\alpha^2_{(j,k)}\leq t}\leq R'^2 t^{\frac{2\gamma}{1+2\gamma}}\right\}.$$

If $\delta\geq\max(\gamma/2,\gamma/(1+2\gamma))$, then the minimax separation rate is

$$\inf_\Phi Sep.Dist^2\simeq L^{-\frac{4\delta}{1+4\delta}}$$

which is achievd by an appropriate choice of the Nested collection of tests up to a $\ln\ln L$.
If $\delta<\gamma/2$ and $\gamma>1/2$, then

$$\inf_\Phi Sep.Dist^2\geq\square\left(\frac{\ln L}{L}\right)^{\frac{2\gamma}{1+2\gamma}},$$

which is achieved by the thresholding test when $\delta\geq\frac{\gamma}{1+2\gamma}$ and which is also the minimax rate of estimation over those spaces. That means that when weak Besov bodies are involved, it is not easier to test than to estimate. More details may be found in [1].

References

[1] M. Fromont, B. Laurent, P. Reynaud-Bouret *Adaptive tests of homogeneity for a Poisson process*, Annales de l'IHP (PS), to appear (2010).

Estimation in Sparse High-Dimensional Trace-Regression

Angelika Rohde
(joint work with Alexandre B. Tsybakov)

Suppose that we observe $(Y_1,X_1),..,(Y_N,X_N)$ related by the *trace-regression model*

$$Y_i=\mathrm{trace}\big(X_i'A^*\big)\ +\ \xi_i,\quad i=1,...,N, \tag{1}$$

where the matrix A^* and the "design" matrices X_i belong to $\mathbb{R}^{m\times T}$, A^* is the unknown parameter of interest and ξ_i are i.i.d. random errors. The emphasis is on estimation of A^* in high-dimensional case, in particular, when the dimension is much greater than the sample size, $mT\gg N$. To make the estimation possible in this setting we assume that A^* is of small rank. Another important remark is that

we mainly focus on very sparse matrices X_i. This means that each X_i contains only a small percentage of non-zero entries. Therefore, multiplication of A^* by X_i masks most of the entries of A^*. Such design matrices will be called *masks*. The following two examples are of particular interest.

(i) *Point masks; matrix completion.* Here $X_i \in \mathcal{X}$ with $\mathcal{X} = \big\{e_k(m)e_l'(T) : 1 \leq k \leqslant m, 1 \leqslant l \leq T\big\}$ and $e_k(m)$ are the canonical basis vectors of $\mathbb{R}^m$. Then each observation Y_i is just one selected entry of A^* corrupted by noise, and the problem is to reconstruct all the entries of A^* (matrix completion). We focus on two special cases of matrix completion, namely, (a) USR (*uniform sampling at random*): X_i are i.i.d. uniformly distributed on $\mathcal{X}$, and (b) *collaborative filtering*, i.e., the trace regression model such that the masks X_i (random or deterministic) belong to $\mathcal{X}$ and are all distinct.

(ii) *Column or row masks.* Each X_i has only one non-zero column or row. This covers the problem known in Machine Learning as multi-task learning. In its simplest version, $N = nT$ with T the number of tasks and n the number of observations per task. The tasks are characterized by vectors of parameters $a_t^* \in \mathbb{R}^m$, $t = 1, \dots, T$, which constitute the columns of matrix A^*: $A^* = (a_1^* \cdots a_T^*)$. In this case the trace regression model (1) can be written as a collection of T standard linear regression models with unknown parameters a_t^*.

We denote by $\|A\|_{S_p}$ the Schatten-p quasi-norm of matrix $A \in \mathbb{R}^{m\times T}$, $0 < p \leqslant \infty$. In order to shrink towards a low-rank representation, we investigate penalized least squares estimators $\hat{A}$ with a Schatten-p penalty term, $0 < p \leq 1$:

$$\hat{A} \in \operatorname*{argmin}_{A\in\mathbb{R}^{m\times T}} \Big\{ \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i - \operatorname{trace}(X_i'A)\Big)^2 + \lambda \|A\|_{S_p}^p \Big\}.$$

We study the convergence of these estimators w.r.t. the Schatten-q quasi-norms and w.r.t. the prediction loss

$$\hat{d}_{2,N}\big(\hat{A}, A^*\big)^2 = \frac{1}{N} \sum_{i=1}^{N} \operatorname{trace}^2\Big(X_i'(\hat{A} - A^*)\Big).$$

We assume below that ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables; for extension to more general ξ_i, see Rohde and Tsybakov (2009). We also assume that $X_1, \dots, X_N$ are independent from $\xi_1, \dots, \xi_N$. We say that the linear map $\mathcal{L} : A \mapsto \big(\operatorname{trace}(X_1'A), ..., \operatorname{trace}(X_N'A)\big)/\sqrt{N}$ (the sampling operator) is uniformly bounded if there exists a constant $c_0 < \infty$ such that $|\mathcal{L}(A)|_2^2 \leq c_0 \|A\|_{S_2}^2$ for all matrices $A \in \mathbb{R}^{m\times T}$ where $|\bullet|_2$ is the Euclidean norm in $\mathbb{R}^N$. An important quantity for our results is the "effective noise level" τ whose values under various assumptions are given in the table below. Here the constants $c > 0, c(p) > 0$ depend only on σ (see Rohde and Tsybakov (2009) for explicit expressions).

Theorem 1. *Let $0 < p \leq 1$, $\lambda = 4\tau$. Then, for cases listed in the table above,*

$$\hat{d}_{2,N}\big(\hat{A}, A^*\big)^2 \leq 16\tau \|A^*\|_{S_p}^p \text{ with probability at least } 1 - \varepsilon,$$

where $\varepsilon = \exp(-C(m + T))$ with a constant $C > 0$ independent of N, m, T.

Assumptions on X_i	Assumptions on N, m, T, p	Value of τ
Uniformly bounded $\mathcal{L}$	$p = 1$	$c((m+T)/N)^{1/2}$
Uniformly bounded $\mathcal{L}$	$0 < p < 1$, $m = T$	$c(p)(m/N)^{1-p/2}$
USR matrix completion	$p = 1$, $(m+T)mT > N$	$c(m+T)/N$
Collaborative filtering	$p = 1$	$c(m+T)^{1/2}/N$

Our second result is valid under a strong condition on the design matrices X_i. We say that the sampling operator $\mathcal{L}$ satisfies the *restricted isometry (RI) condition* $\mathrm{RI}(r, \nu)$ for some integer $1 \le r \le \min(m, T)$ and some $0 < \nu < \infty$ if there exists a constant $\delta_r \in (0, 1)$ such that

$$(1 - \delta_r)\|A\|_{S_2} \;\le\; \nu|\mathcal{L}(A)|_2 \;\le\; (1 + \delta_r)\|A\|_{S_2}$$

for all matrices $A \in \mathbb{R}^{m \times T}$ of rank at most r. This differs from the RI condition introduced by Candes and Tao (2005) in the vector case or from its analog for the matrix case suggested by Recht et al. (2007) because we have here the scaling factor $\nu \neq 1$. It accounts for the fact that the masks X_i can be very sparse, so that they do not induce isometries with coefficient close to one.

Theorem 2. *Let $0 < p \le 1$ and $\mathrm{rank}(A^*) \leqslant r$. Assume that condition RI $((2+a)r)$,ν) holds with some $0 < \nu < \infty$, with a sufficiently large $a = a(p)$ depending only on p and with $0 < \delta_{(2+a)r} \leqslant \delta_0$ for a sufficiently small $\delta_0 = \delta_0(p)$ depending only on p. Then, for $\lambda = 4\tau$ with τ as in the first two lines of the table above we have*

$$\begin{aligned} \hat{d}^2_{2,N}\big(\hat{A}, A^*\big) \;&\le\; C_1 r\, \tau^{\frac{2}{2-p}} \nu^{\frac{2p}{2-p}}, \\ \|\hat{A} - A^*\|^q_{S_q} \;&\le\; C_2 r\, \tau^{\frac{q}{2-p}} \nu^{\frac{2q}{2-p}}, \quad \forall\, q \in [p, 2], \end{aligned} \tag{2}$$

with probability at least $1 - \varepsilon$, where ε is as in Theorem 1, and $C_1, C_2 > 0$ are constants such that C_1 depends only on p and C_2 depends on p and q.

On the difference from Theorem 1, here we have bounds not only for the prediction loss but also for direct estimation of A^*, cf. (2). However, the restricted isometry is not suitable for sparse design matrices as in examples (i) and (ii) above, since there ν is large. For the multi-task learning model $\nu \sim \sqrt{T}$, while for USR we have $\nu = \sqrt{mT}$. What is more, for USR, the RI condition can be satisfied only if $N > mT$, which is in contradiction with the matrix completion context. Nevertheless, we can still use Theorem 1 where the bound depends on the magnitude of singular values of A^*. This dependence is weaker for smaller p. which leads us to considering $p = p(N, m, T)$ that is small for large N. We show that, under the assumptions of the second line in the table above and provided the singular values of A^* are not exponentially large, the estimator $\hat{A}$ with $\lambda = 4\tau$ and $p = (\log(N/m))^{-1}$ satisfies, with probability at least $1 - \varepsilon$,

$$\hat{d}^2_{2,N}\big(\hat{A}, A^*\big) \;\le\; C\frac{rm}{N} \log\Big(\frac{N}{m}\Big), \quad \text{where } r = \mathrm{rank}(A^*), \tag{3}$$

ε is as in Theorem 1 and constant $C > 0$ depends only on σ.

The proofs are based on tools from the theory of empirical processes. As a by-product we derive bounds for the kth entropy numbers of the quasi-convex Schatten class embeddings $S_p^m \hookrightarrow S_2^m$, $p < 1$, which are of independent interest.

Optimality issues. The singular value decomposition reveals that an $m \times m$-matrix A of $\mathrm{rank}(A) = r$ has $(2m-r)r$ degrees of freedom, i.e., is characterized by this number of parameters (effective dimension). Under some regularity conditions on the masks X_i one can show that the optimal rate of estimation under the loss $\hat{d}^2_{2,N}(\bullet, \bullet)$ has the form "(effective dimension)/sample size". In case $r \ll m$ this is of order rm/N, which coincides with the bound in (3) up to a log factor. Under the RI condition with $\nu = 1$, this lower bound for estimation in the Frobenius norm has recently been proved by Candès and Plan (2010). Note that imposing this condition (which is also the case in the work of Negahban and Wainwright (2009) parallel to ours) means focusing on "full" (non-sparse) matrices X_i, such as matrices with all entries being i.i.d. Rademacher or Gaussian random variables. These two papers prove results analogous to (2) with $\nu = 1$, $p = 1$, $q = 2$.

References

[1] E. J. Candès and Y. Plan, *Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements*, `arxiv:1001.0339`

[2] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory **51** (2005), 4203–4215.

[3] S. Negahban and M. J. Wainwright, *Estimation of (near) low-rank matrices with noise and high-dimensional scaling*, `arxiv:0912.5100`

[4] B. Recht, M. Fazel and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, `arxiv:0706.4138`

[5] A. Rohde and A. B. Tsybakov, *Estimation of high-dimensional low rank matrices*, `arxiv:0912.5338`

Optimal rates of sparse estimation and universal aggregation

Alexandre Tsybakov

(joint work with Philippe Rigollet)

Let $\mathcal{Z} = \{(x_1, Y_1), \dots, (x_n, Y_n)\}$ be a collection of independent random couples such that $(x_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ is an arbitrary set. We consider the regression model

$$Y_i = \eta(x_i) + \xi_i, \ i = 1, \dots, n\,,$$

where $\eta : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and the errors ξ_i are independent Gaussian $\mathcal{N}(0, \sigma^2)$. The covariates are deterministic elements $x_1, \dots, x_n$ of $\mathcal{X}$. For any function $f : \mathcal{X} \to \mathbb{R}$ define $\|f\|^2 = n^{-1} \sum_{i=1}^n f^2(x_i)$.

Given a dictionary $\mathcal{H} = \{f_1, \dots, f_M\}$ of functions $f_j : \mathcal{X} \to \mathbb{R}$ such that $\max_{1 \leqslant j \leqslant M} \|f_j\| \leqslant 1$, and a subset $\Theta \subseteq \mathbb{R}^M$, the goal of *aggregation* is to find an estimator $\hat{\eta}$ that mimics the best linear combination $\mathsf{f}_\vartheta = \sum_{j=1}^M \vartheta_j f_j, \vartheta \in \Theta$, in the sense that the excess risk defined by

$$\mathbb{E}\|\hat{\eta} - \eta\|^2 - \min_{\vartheta \in \Theta} \|\mathsf{f}_\vartheta - \eta\|^2$$

is as small as possible. For different choices of Θ this problem has been studied by Nemirovski [3], Tsybakov [5], Bunea et al. [1], Lounici [2]; the performance of $\hat{\eta}$ is usually assessed via an oracle inequality of the form

$$\mathbb{E}\|\hat{\eta}-\eta\|^2 \leqslant (1+\varepsilon)\min_{\vartheta\in\Theta}\|\mathsf{f}_\vartheta-\eta\|^2+\Delta_{n,M}(\Theta)\,,\quad \epsilon\geqslant 0\,. \tag{1}$$

The smallest possible (in a minimax sense) remainder term $\Delta_{n,M}(\Theta)$ characterizes the price for aggregation over Θ and is called *optimal rate of aggregation on* Θ, using the terminology in Tsybakov [5]. Here we focus on exact oracle inequalities, i.e., those with $\epsilon=0$ allowing for meaningful bounds on the excess risk. The best approximation is obtained by choosing $\Theta=\mathbb{R}^M$. In this case it can be shown that the smallest possible remainder term is large, $\Delta_{n,M}(\mathbb{R}^M)=CM/n, C>0$, and that it is attained by a simple least squares estimator.

We propose an estimator $\hat{\eta}=\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}}$, that satisfies a stronger type of oracle inequality than (1). In particular it can still give an informative result when M is much larger than n by adapting to the underlying sparsity of the problem.

Let $|\bullet|_1$ denote the ℓ_1 norm in $\mathbb{R}^M$ and $M(\vartheta)$ denote the ℓ_0 norm of $\vartheta\in\mathbb{R}^M$, i.e., the number of non-zero coordinates of ϑ. Set $B_i(r)=\{\vartheta\in\mathbb{R}^M : |\vartheta|_i\leqslant r\}$, $r>0$, $i=0,1$. For two real numbers a and b define $a\vee b=\max(a,b)$ and $a\wedge b=\min(a,b)$. We denote by $\mathbf{X}$ the $n\times M$ design matrix with elements $\mathbf{X}_{i,j}=f_j(x_i)$ and rank $R=\mathrm{rank}(\mathbf{X})\leqslant M\wedge n$. Finally, for any $\vartheta\in\mathbb{R}^M$, define $\widetilde{M}(\vartheta)=M(\vartheta)\wedge R$.

We call the *sparsity pattern* a binary vector $\mathsf{p}\in\mathcal{P}=\{0,1\}^M$ and define $|\mathsf{p}|=M(\mathsf{p})$ the number of ones in p. Let $\mathbb{R}^{\mathsf{p}}\subseteq\mathbb{R}^M$ defined by $\mathbb{R}^{\mathsf{p}}=\{\vartheta\bullet\mathsf{p} : \vartheta\in\mathbb{R}^M\}$, where $\vartheta\bullet\mathsf{p}\in\mathbb{R}^M$ is the vector with coordinates $(\vartheta\bullet\mathsf{p})_j=\vartheta_j\mathsf{p}_j, j=1\ldots,M$. Then, for any $\mathsf{p}\in\mathcal{P}$, we can define a least squares estimator $\hat{\vartheta}_{\mathsf{p}}$ on $\mathbb{R}^{\mathsf{p}}$ by

$$\hat{\vartheta}_{\mathsf{p}}\in\operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^{\mathsf{p}}}|\mathbf{Y}-\mathbf{X}\vartheta|_2^2\,.$$

Let $\nu=(\nu_{\mathsf{p}})_{\mathsf{p}}$ be a probability measure on $\mathcal{P}$ defined by $\nu_{\mathsf{p}}=0$ if $|\mathsf{p}|>R$ and by,

$$\nu_{\mathsf{p}}\propto\exp\Big(-\frac{1}{4\sigma^2}\sum_{i=1}^n(Y_i-\mathsf{f}_{\hat{\vartheta}_{\mathsf{p}}}(x_i))^2-\frac{|\mathsf{p}|}{2}\Big)\left(\frac{|\mathsf{p}|}{2eM}\right)^{|\mathsf{p}|}\,,\ \text{if } |\mathsf{p}|\leqslant R.$$

Notice that second factor in the definition of ν_{p} exponentially downweights sparsity patterns p with large $|\mathsf{p}|$, i.e., that are not sparse.. The *Exponential Screening (ES)* aggregate is defined as the linear combination $\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}}$ where

$$\tilde{\vartheta}^{\mathrm{ES}}=\sum_{\mathsf{p}\in\mathcal{P}}\hat{\vartheta}_{\mathsf{p}}\nu_{\mathsf{p}}\,.$$

A tractable numerical approximation of $\tilde{\vartheta}^{\mathrm{ES}}$ using the Metropolis-Hastings algorithm is detailed in [4].

Theorem 1. *For any $M\geqslant 1, n\geqslant 1$, the ES aggregate $\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}}$ satisfies*

$$\mathbb{E}\|\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}}-\eta\|^2\leqslant\min_{\vartheta\in\mathbb{R}^M}\left\{\|\mathsf{f}_\vartheta-\eta\|^2+\varphi_{n,M}(\vartheta)\right\}+\frac{\sigma^2}{n}(5\log(1+eM)+8\log 2). \tag{2}$$

where $\varphi_{n,M}(0) = 0$ *and, for* $\vartheta \neq 0$,

$$\varphi_{n,M}(\vartheta) = \min\left\{ \frac{5\sigma^2 \widetilde{M}(\vartheta)}{n} \log\left(\frac{eM}{\widetilde{M}(\vartheta) \vee 1}\right), \frac{8\sigma|\vartheta|_1}{\sqrt{n}} \sqrt{\log\left(1 + \frac{3eM\sigma}{|\vartheta|_1\sqrt{n}}\right)} \right\} .$$

Furthermore, if there exists $\vartheta^* \in \mathbb{R}^M$ *such that* $\eta = \mathsf{f}_{\vartheta^*}$, *we have*

$$\mathbb{E}\|\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}} - \mathsf{f}_{\vartheta^*}\|^2 \leqslant \psi_{n,M}(\vartheta^*) + \frac{8\sigma^2}{n} \log 2 .$$

where $\psi_{n,M}(0) = 0$ *and, for* $\vartheta \neq 0$,

$$\psi_{n,M}(\vartheta) = \min\left\{ \frac{5\sigma^2 \widetilde{M}(\vartheta)}{n} \log\left(\frac{eM}{\widetilde{M}(\vartheta) \vee 1}\right), \frac{8\sigma|\vartheta|_1}{\sqrt{n}} \sqrt{\log\left(1 + \frac{3eM\sigma}{|\vartheta|_1\sqrt{n}}\right)}, 4|\vartheta|_1^2 \right\} .$$

Moreover, we show that the rate $\psi_{n,M}(\vartheta)$ is optimal in a minimax sense on the intersection of ℓ_0 and ℓ_1 balls. Note that the sparsity oracle inequality (2) is much stronger than oracle inequalities such as (1). Indeed, if there exists a sparse vector $\bar{\vartheta}$ such that both the approximation term $\|\mathsf{f}_{\bar{\vartheta}} - \eta\|^2$ and the stochastic error term $\varphi_{n,M}(\bar{\vartheta})$ are small, then the ES aggregate adapts to it. Furthermore, this result captures three measures of sparsity: the ℓ_0 norm $M(\vartheta)$, the ℓ_1 norm $|\vartheta|_1$ and the rank R.

For a given $\Theta \subseteq \mathbb{R}^M$, the goal of aggregation is to find an estimator $\hat{\eta}$ (possibly depending on Θ) that satisfies (1), ideally with $\epsilon = 0$. Five choices for Θ have been proposed and studied in the literature. They are summarized in the accompanying table, where D is an integer between 1 and M. Bunea et al. [1] raised the issue of

Problem	Θ	Description
(MS)	$\Theta_{(\mathrm{MS})} = B_0(1) \cap B_1(1)$	Best in dictionary
(C)	$\Theta_{(\mathrm{C})} = B_1(1)$	Best convex combination
(L)	$\Theta_{(\mathrm{L})} = \mathbb{R}^M = B_0(M)$	Best linear combination
(L_D)	$\Theta_{(\mathrm{L}_D)} = B_0(D)$	Best D-sparse linear comb.
(C_D)	$\Theta_{(\mathrm{C}_D)} = B_0(D) \cap B_1(1)$	Best D-sparse convex comb.

universal aggregation, i.e., of finding estimators that solve all aggregation problems at once. They proved that the BIC estimator, which does not depend on Θ, solves the first four problems in the table above in an approximate sense. The next theorem shows that the ES aggregate fully realizes universal aggregation, i.e., solves simultaneously all five problems with remainder terms $\Delta^*_{n,M}(\Theta)$, which we prove to be optimal rates of aggregation in a minimax sense.

Theorem 2. *For any* $M \geqslant 2, n \geqslant 1, D \leqslant M$, *and for either of the five sets* Θ *in the table above the ES aggregate* $\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}}$ *satisfies the following oracle inequality*

$$\mathbb{E}\|\mathsf{f}_{\tilde{\vartheta}^{\mathrm{ES}}} - \eta\|^2 \leqslant \min_{\vartheta \in \Theta} \|\mathsf{f}_\vartheta - \eta\|^2 + C\Delta^*_{n,M}(\Theta) ,$$

where $C > 0$ is a numerical constant and

$$\Delta^*_{n,M}(\Theta) = \begin{cases} \frac{\sigma^2 \log M}{n} & \text{if } \Theta = \Theta_{(\mathrm{MS})}, \\ \sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2(M\wedge R)}{n} \log\left(1 + \frac{eM}{M\wedge R}\right) & \text{if } \Theta = \Theta_{(\mathrm{C})}, \\ \frac{\sigma^2(M\wedge R)}{n} \log\left(1 + \frac{eM}{M\wedge R}\right) & \text{if } \Theta = \Theta_{(\mathrm{L})}, \\ \frac{\sigma^2(D\wedge R)}{n} \log\left(1 + \frac{eM}{D\wedge R}\right) & \text{if } \Theta = \Theta_{(\mathrm{L}_D)}, \\ \sqrt{\frac{\sigma^2}{n} \log\left(1 + \frac{eM\sigma}{\sqrt{n}}\right)} \wedge \frac{\sigma^2(D\wedge R)}{n} \log\left(1 + \frac{eM}{D\wedge R}\right) & \text{if } \Theta = \Theta_{(\mathrm{C}_D)}. \end{cases}$$

This theorem improves upon the previously known results for two reasons: universal aggregation is solved by ES with $\epsilon = 0$ and the rates can be better if R is small.

References

[1] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, *Aggregation for Gaussian regression*, Ann. Statist., **35** (2007): 1674–1697.

[2] K. Lounici, *Generalized mirror averaging and D-convex aggregation*, Math. Methods Statist. **16** (2007): 246–259.

[3] A. Nemirovski, *Topics in non-parametric statistics*, In Lectures on probability theory and statistics (Saint-Flour, 1998), vol. **1738** of Lecture Notes in Math., Springer, Berlin, 85–277

[4] P. Rigollet, and A. Tsybakov, *Exponential Screening and optimal rates of sparse estimation*, arXiv:1003.2654 (2010).

[5] A. B. Tsybakov, *Optimal rates of aggregation*, In *COLT* (B. Schölkopf and M. K. Warmuth, eds.), vol. **2777** (2003) of Lecture Notes in Computer Science, Springer, 303–313.

Bayesian Regularization

Aad van der Vaart

The last decades have seen a growing interest in Bayesian methods for recovering curves, surfaces or other high-dimensional objects from noisy measurements. The object ϑ is modelled as a realization from some *prior* probability distribution Π, and the observed data X is viewed as drawn from a probability density $x \mapsto p_\vartheta(x)$ that depends on the realization of ϑ. The *posterior* distribution of the "parameter" ϑ is then given by Bayes' rule as

$$d\Pi(\vartheta|\, X) \propto p_\vartheta(X)\, d\Pi(\vartheta).$$

To investigate the quality of the posterior distribution we put ourselves in a non-Bayesian framework, where it is assumed that the data X are generated according to the density p_{ϑ_0} determined by a fixed parameter ϑ_0, and view the posterior distribution as just a random measure on the parameter space. The Bayesian procedure is considered accurate if this random measure concentrates its mass near the parameter ϑ_0. We wish this to be true for many ϑ_0 simultaneously, preferably uniformly in ϑ_0 belonging to a class of test models. For instance, a set of surfaces known to have a certain number of bounded derivatives.

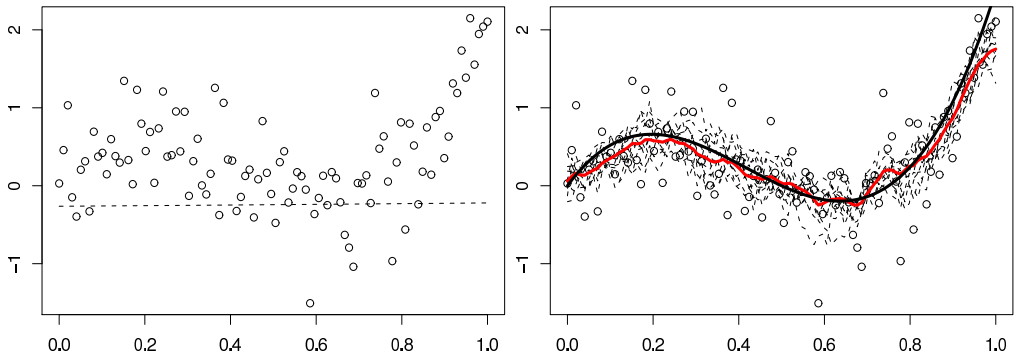

FIGURE 1. Observations in a regression problem. One realization from a Gaussian process prior (left panel) and 10 from the posterior distribution (right panel). The true regression curve and the posterior mean are indicated in the right panel. The Bayesian updating is successful: the realizations from the posterior are much closer to the truth than those from the prior.

Except in very special cases this question can be investigated only in an asymptotic setting. We consider data X^n depending on an index n (for instance sample size) and study the resulting sequence of posterior distributions $d\Pi_n(\vartheta|\, X^n)$ as $n \to \infty$. In a setting where the informativeness of the data increases indefinitely with n, we desire that this sequence contracts to the Dirac measure at ϑ_0, meaning complete recovery "in the limit". Given a metric structure d on the parameter space, we can more precisely measure the *rate of contraction*. We say that this is at least ε_n if, for any sequence of constants $M_n \to \infty$,

$$\Pi_n\big(\vartheta : d(\vartheta, \vartheta_0) < M_n\varepsilon_n|\, X^n\big) \to 1.$$

The convergence can be in mean, or in the almost sure sense. Thus the posterior distribution puts almost all its mass on balls of radius of the order ε_n around ϑ_0.

In classical finite-dimensional problems, with ϑ a vector in Euclidean space and n the sample size, the rate of contraction ε_n is typically $n^{-1/2}$, relative to for instance the Hellinger distance, for any prior with full support. The *Bernstein-Von Mises theorem* makes this preciser in a normal approximation to the posterior distribution. The prior distribution does not appear in this approximation and is said to "wash out" as $n \to \infty$. In nonparametric problems this is very different. First there are many priors which do not lead to contraction of the posterior at all. Second many natural priors yield a rate of contraction that depends on the combination of the prior and the true parameter. The positive news is that a good match between prior and ϑ_0 may lead to an *optimal rate of contraction*, equal to the minimax rate for a problem.

In practice such "good matches" may not be easy to achieve. It is never trivial to have a proper intuitive understanding of a prior probability distribution on an infinite-dimensional set. Furthermore, and more importantly, one does not know

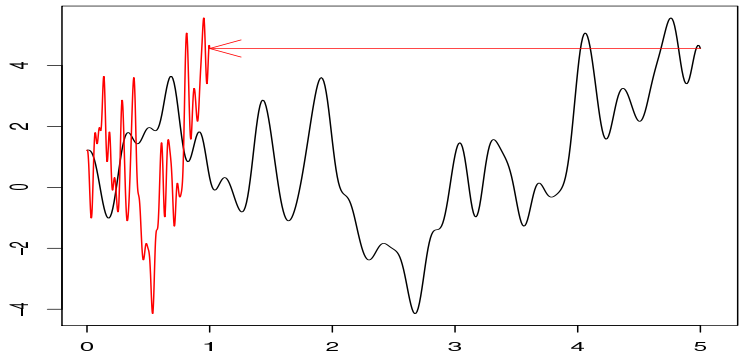

FIGURE 2. A realization of the squared exponential processes and its rescaling to the unit interval.

the fine properties of the true parameter ϑ_0. The elegant solution to the dilemma of prior choice (a classical point of criticism to Bayesian methods) is to work with many priors at the same time. We start with a collection of priors Π_α, indexed by some parameter α in an index set A, which is assumed to contain at least one appropriate prior for each possible truth ϑ_0. Next we combine these priors by putting a prior distribution, a *hyper prior*, on the index α. If A is countable and the hyper prior is denoted by $(l_\alpha : \alpha \in A)$, then this this just leads to the overall prior

$$\Pi = \sum_\alpha l_\alpha \Pi_\alpha.$$

Inference, using Bayes' rule, proceeds as before. The hope is that the data will automatically "use" the priors Π_α that are appropriate for ϑ_0, and produce a posterior that contracts at an optimal rate, given that at least one of the priors Π_α would produce this rate if used on its own.

This automatic *adaptation* of the posterior distribution sounds too good to be true. Obviously, it depends strongly on the weights $l = (l_\alpha : \alpha \in A)$ and the prior distributions Π_α. Because the latter often possess very different "dimensionalities", finding appropriate weights can be delicate. However, quite natural schemes turn out to do the job, although sometimes a logarithmic factor is lost. In our talk we first presented an abstract result, which shows that *adaptation is easy*, in the sense that many weight distributions l work. Next we considered adaptation to the regularity of a surface ϑ_0 using Gaussian process priors. We showed that the "Bayesian" regularity of the sample paths can be changed by stretching or shrinking them. By scaling a process with analytic sample paths (for instance the squared exponential process) with a random variable, it becomes a suitable prior for truths of any regularity. Finally we considered adaptation to sparsity in the many normal means problem, where the parameter (the mean vector) is a vector in $\mathbb{R}^n$, many of whose coordinates are thought to be zero. By putting a prior on the number of nonzero means, next and independent heavy-tailed (Laplace

or polynomial-tailed) on a randomly chosen subset of nonzero parameters, it is possible to adapt to spaces of sparse parameters.

Details on these results can be found in the papers [2], [3] and [1].

References

[1] I. Castillo and A.W. van der Vaart, *Needles and straws in a haystack: posterior concentration for possibly sparse sequences*, Preprint (2010).

[2] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart, *Nonparametric Bayesian model selection and averaging*, Electron. J. Stat. **2** (2008): 63–89.

[3] A. W. van der Vaart and J. H. van Zanten. *Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth*, Ann. Statist. **37(5B)** (2009): 2655–2675.

Statistical recovery in high dimensions: A unified analysis of decomposable regularizers

Martin Wainwright

(joint work with Sahand Negahban, Pradeep Ravikumar and Bin Yu)

There are a variety of results on the behavior of different M-estimators under high-dimensional scaling, including ℓ_1-regularized linear regression, block-norm regularization for multivariate problems, and nuclear-norm regularizer for low-rank matrices. In this talk, we present a single theorem that isolates some properties common to many high-dimensional analyses, and yields optimal rates for a class of regularized M-estimators. The result depends on two intuitive conditions: the regularizer needs suitably constrain the parameter space via the notion of decomposability, and the loss function needs to be sufficiently curved, formalized via the notion of restricted strong convexity. When applied to specific high-dimensional models, we recover various results (some known and some new), including rates for regression models under both weak and hard sparsity constraints, estimating sparse (generalized) linear models, structured covariance matrices, and near low-rank matrices. In many cases, the upper bounds are matched (up to constant factors) by minimax lower bounds.

References

[1] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,* Proceedings of the NIPS Conference, December 2009.

Adaptive Threshold estimation: Minimaxity, Empirical Bayes and Loss Minimization

Cun-Hui Zhang

1. Summary. We discuss a number of optimality criteria in the normal mean problem with the ℓ_2 loss, including adaptive minimaxity, general empirical Bayes, and minimum risk and loss for threshold estimation. A general maximum likelihood empirical Bayes method is shown to possess adaptive ratio optimality and adaptive minimax properties via an oracle inequality. The FDR threshold level is shown to nearly minimize the risk and loss among all soft threshold estimators. An oracle inequality is provided for the adaptive soft-threshold estimator for dependent data.

2. Adaptive minimaxity. Let

$$\boldsymbol{X} \sim N(\boldsymbol{\vartheta}, \boldsymbol{\Sigma}), \quad X_i \sim N(\vartheta_i, 1).$$

We consider the estimation of $\boldsymbol{\vartheta} \in \mathbb{R}^n$ with data $\boldsymbol{X}$ under the ℓ_2 loss. For $\Theta \subseteq \mathbb{R}^n$, the minimax risk is defined as

$$\mathscr{R}_n(\Theta) = \inf_{all\ \boldsymbol{\delta}} \sup_{\boldsymbol{\vartheta}\in\Theta} E\|\boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n.$$

Let $\|\boldsymbol{\vartheta}\|_p = (\sum_{i=1}^n |\vartheta_i|^p)^{1/p}$ for $p > 0$ and $\|\boldsymbol{\vartheta}\|_0 = \#\{i \leqslant n : \vartheta_i \neq 0\}$. The ℓ_p balls are defined as $\Theta_{p,C,n} = \{\boldsymbol{\vartheta} : \|\boldsymbol{\vartheta}\|_p^p/n \leqslant C^p\}$ for $p > 0$ and $\Theta_{0,C,n} = \{\boldsymbol{\vartheta} : \|\boldsymbol{\vartheta}\|_0/n \leqslant C\}$. Let $\mathscr{T}_s$ and $\mathscr{T}_h$ be respectively the classes of soft and hard threshold estimators. For $0 \leqslant p \leqslant 2$, the minimax risk in small ℓ_p balls can be approximated by the minimax risk of threshold estimators:

$$\inf_{\mathscr{T}} \sup_{\boldsymbol{\vartheta}\in\Theta_{p,C_n,n}} E\|\boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n = (1+o(1))\mathscr{R}_n(\Theta_{p,C_n,n}),$$

where $\mathscr{T} = \mathscr{T}_s$ or $\mathscr{T} = \mathscr{T}_h$, provided $0 < \mathscr{R}_n(\Theta_{p,C_n,n}) \ll 1$. This result was proved earlier by Donoho and Johnstone (1994) under the condition $(\log n)/n \ll \mathscr{R}_n(\Theta_{p,C_n,n}) \ll 1$.

A sequence of estimators $\{\boldsymbol{\delta}_n, n \geqslant 1\}$ is adaptive minimax if

$$\sup_{\boldsymbol{\vartheta}\in\Theta_n} E\|\boldsymbol{\delta}_n(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n = (1+o(1))\mathscr{R}_n(\Theta_n)$$

for many (large and small) sequences parameter classes $\{\Theta_n\}$, $\Theta_n \subseteq \mathbb{R}^n$.

3. Empirical Bayes. The general empirical Bayes benchmark is

$$R_n^*(\boldsymbol{\vartheta}) = \inf_{t(\bullet)} E\|t(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n,$$

where the infimum is taken over all univariate Borel functions (Robbins, 1951, 1956). This benchmark is attained with the Bayes rule

$$R_n^*(\boldsymbol{\vartheta}) = E\|t_{G_n}^*(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n,\ t_G^*(x) = \frac{\int \vartheta f(x|\vartheta)G(d\vartheta)}{\int f(x|\vartheta)G(d\vartheta)}$$

where $G_n(d\vartheta) = \sum_{i=1}^n I\{\vartheta_i \in d\vartheta\}/n$ and $f(x|\vartheta)$ is the $N(\vartheta, 1)$ density.

In empirical Bayes, the performance of an estimator is typically measured through an oracle inequality, or equivalently an upper bound on

$$\text{the regret } = E\|\boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n - R_n^*(\boldsymbol{\vartheta}).$$

The empirical Bayes benchmark focuses on the risk at the "true" $\boldsymbol{\vartheta}$. Greenshtein and Ritov (2009) proved

$$R_n^*(\boldsymbol{\vartheta}) = (1+o(1)) \inf_{\boldsymbol{\delta}\in\mathscr{E}} E\|\boldsymbol{\delta}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n,$$

where $\mathscr{E}$ is the class of all estimators satisfying $\boldsymbol{\delta}(T(\boldsymbol{X})) = T(\boldsymbol{\delta}(\boldsymbol{X}))$ for all permutations T. Empirical Bayes is connected to the minimax criterion via

$$\max_{\boldsymbol{\vartheta}\in\Theta_{p,C,n}} R_n^*(\boldsymbol{\vartheta}) = (1+o(1))\mathscr{R}_n(\Theta_{p,C,n}).$$

The general empirical Bayes approach requires nonparametric estimation of the oracle Bayes rule $t^*_{G_n}$. Parametric Bayes approaches include the James-Stein (1961) estimator (Efron and Morris, 1972, 1973) and the EBThresh of Johnstone and Silverman (2004).

4. General maximum likelihood empirical Bayes. Jiang and Zhang (2009) proposed the following general maximum likelihood empirical Bayes (GMLEB) estimator:

$$\widehat{\boldsymbol{\vartheta}} = t^*_{\widehat{G}_n}(\boldsymbol{X}),\ \widehat{G}_n = \arg\max_G \prod_{i=1}^n \int f(X_i|\vartheta)G(d\vartheta).$$

For $\boldsymbol{\Sigma} = \boldsymbol{I}$, they proved the following oracle inequality:

$$E\|t^*_{\widehat{G}_n}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n \leqslant \left(\sqrt{R^*(\boldsymbol{\vartheta})} + M_0\sqrt{\epsilon(n,G_n,p)}\right)^2$$

for all $\boldsymbol{\vartheta}\in\mathbb{R}^n$ and $2/\log n \leqslant p \leqslant \infty$, where M_0 is a universal constant and

$$\epsilon(n,G,p) = \max\left\{\frac{2\log n}{n}, \left[\frac{\sqrt{\log n}\mu_p^w(G)}{n}\right]^{p/(1+p)}\right\}(\log n)^4,$$

with $\mu_p^w(G) = \{\sup_x |x|^p G([-x,x]^c)\}^{1/p}$ being the weak p-norm of G. It follows from this oracle inequality that the general empirical Bayes benchmark is approximately achieved when
$\inf_{p>2/\log n}\epsilon(n,G_n,p) \ll R^*(\boldsymbol{\vartheta})$. Moreover, the oracle inequality implies the adaptive minimaxity of the GMLEB in ℓ_p balls $\Theta_{p,C_n,n}$ within the range

$$(\log n)^{4+p/2+3/p}/n \ll C_n^p \ll n^p/(\log n)^{4+9p/2}.$$

These theoretical results indicate that the GMLEB should perform well for sparse and dense $\boldsymbol{\vartheta}$ in general, except for extremely sparse $\boldsymbol{\vartheta}$. Our simulation results support this claim.

5. Adaptive threshold estimation. The logarithmic factor in the lower bound of C_n^p for the adaptive minimxity of the GMLEB in ℓ_p balls leads to our consideration of adaptive thresholding methods for the estimation of a (possibly extremely) sparse $\boldsymbol{\vartheta}$.

For $\boldsymbol{\Sigma} = \boldsymbol{I}$, Abramovich, Benjamini, Donoho and Johnstone (2006) proved that at the FDR threshold level with a nominal FDR level $q_0 \in (0, 1/2]$, the

hard threshold estimator is adaptive minimax in ℓ_p balls when $(\log n)^{6-p/2}/n \ll \mathscr{R}(\Theta_{p,C_n,n}) \ll n^{-\delta_0}$, $\delta_0 > 0$, but the same method with $q_0 \in (1/2, 1)$ is not adaptive minimax in such ℓ_p balls (off by a constant factor). Johnstone and Silverman (2004) proved the adaptive rate minimaxity of the EBThresh when $(\log n)^{3-p/2}/n \ll \mathscr{R}(\Theta_{p,C_n,n}) \ll 1$.

We propose to use the soft threshold estimator at the FDR threshold level:

$$\widehat{\boldsymbol{\vartheta}} = s_{\widehat{\lambda}}(\boldsymbol{X}), \quad \widehat{\lambda} = \min\big\{k : N(\xi_k) \geqslant k\big\},$$

where $s_\lambda(x) = \mathrm{sgn}(x)(|x| - \lambda)_+$, $N(t) = \#\{i \leqslant n : |X_i| > t\}$ and $2\Phi(-\xi_k) = q_0 k/n$ with the $N(0,1)$ distribution function $\Phi(\bullet)$.

We prove the following oracle inequalities for this adaptive soft threshold estimator:

$$E\|s_{\widehat{\lambda}}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n \leqslant \Big(\sqrt{R_n^{soft}(\boldsymbol{\vartheta})} + M_0\sqrt{\epsilon_n^{soft}(\boldsymbol{\vartheta})}\Big)^2, \ \forall\ \boldsymbol{\vartheta} \in \mathbb{R}^n,$$

where $\epsilon_n^{soft}(\boldsymbol{\vartheta}) = \|\boldsymbol{\Sigma}\|(\log n)^\kappa/n + \sum_{i=1}^n (|\vartheta_i|^2 \wedge 1)/n$ with $\kappa = I\{\boldsymbol{\Sigma} \neq \boldsymbol{I}\}$ and the spectrum norm $\|\boldsymbol{\Sigma}\|$, and

$$R_n^{soft}(\boldsymbol{\vartheta}) = \min_{\lambda > 0}\Big\{2\Phi(-\lambda) + \sum_{i=1}^n \frac{\vartheta_i^2 \wedge (\lambda^2 + 1)}{n}\Big\}.$$

For $\boldsymbol{\Sigma} = \boldsymbol{I}$, this oracle inequality implies the adaptive minimaxity of the estimator in all ℓ_p balls satisfying $1/n \ll \mathscr{R}_n(\Theta_{p,C_n,n}) \ll 1$. Moreover, for dependent data, we proved that the adaptive soft threshold estimator approximately achieves loss minimization in the sense of

$$E\|s_{\widehat{\lambda}}(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n \leqslant (1 + o(1))E\inf_\lambda \|s_\lambda(\boldsymbol{X}) - \boldsymbol{\vartheta}\|^2/n + M_0\epsilon_n^{soft}(\boldsymbol{\vartheta}),$$

for all $\boldsymbol{\vartheta} \in \mathbb{R}^n$.

6. A final remark. Similar to the empirical Bayes and adaptive threshold approaches where parameters of estimators (G_n and λ) are replaced by suitable estimates, aggregation methods (Rigollet and Tsybakov, 2010) take a convex or other types of combination of estimators within a certain class. It is unclear if such aggregation methods possesses similar or stronger oracle properties than the maximum likelihood empirical Bayes and adaptive soft threshold estimators.

References

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

[2] D.L. Donoho, , and I.M. Johnstone, *Minimax risk over ℓ_p-balls for ℓ_q-error*, Probab. Theory Related Fields **99** (1994), 277–303.

[3] B. Efron, and C. Morris, *Empirical Bayes on vector observations: An extension of Stein's method*, Biometrika **59** (1972), 335–347.

[4] B. Efron, and C. Morris, *Stein's estimation rule and its competitors–an empirical Bayes approach*, J. Amer. Statist. Assoc. **68** (1973), 117–130.

[5] E. Greenshtein, and Y. Ritov, *Asymptotic efficiency of simple decisions for the compound decision problem*, In *Optimality: The Third Erich L. Lehmann Symposium* J. Rojo, Ed. Institute of Mathematical Statistics, Lecture Notes-Monograph Series **57** (2009), 266–275

[6] W. James, and C. Stein, *Estimation with quadratic loss*, Proc. Fourth Berkeley Symp. Math. Statist. Probab. **1** (1961), 361-379.
[7] W. Jiang, and C.H. Zhang, *General maximum likelihood empirical Bayes estimation of normal means*, Ann. Statist. **37** (2009), 1647–1684.
[8] I.M. Johnstone, and B.W. Silverman, *Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences*, Ann. Statist. **32** (2004): 1594–1649.
[9] H. Robbins, *Asymptotically subminimax solutions of compound statistical decision problems*, Proc. Second Berkeley Symp. Math. Statist. Probab. **1** (1951), 131–148.
[10] H. Robbins, *An empirical Bayes approach to statistics*, Proc. Third Berkeley Symp. Math. Statist. Probab. **1** (1956), 157–163.
[11] P. Rigollet, and A.B. Tsybakov, *Exponential screening and optimal rates of sparse estimation*, arXiv:1003.2654v2 (2010).

Optimal Estimation of Large Covariance Matrices

Harrison H. Zhou
(joint work with T. Tony Cai, Cun-Hui Zhang)

Suppose we observe independent and identically distributed p-variate random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with covariance matrix $\Sigma_{p\times p}$ and the goal is to estimate the unknown matrix $\Sigma_{p\times p}$ based on the sample $\{\mathbf{X}_i : i = 1, ..., n\}$. This covariance matrix estimation problem is of fundamental importance in multivariate analysis. A wide range of statistical methodologies, including clustering analysis, principal component analysis, linear and quadratic discriminant analysis, regression analysis, require the estimation of the covariance matrices. With dramatic advances in technology, large high-dimensional data are now routinely collected in scientific investigations. Examples include climate studies, gene expression arrays, functional magnetic resonance imaging, risk management and portfolio allocation and web search problems. In such settings, the standard and most natural estimator, the sample covariance matrix, often performs poorly. Regularization methods, originally developed in nonparametric function estimation, have recently been applied to estimate large covariance matrices.

Following Bickel and Levina (2008a) we consider estimating the covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{1\le i,j\le p}$ over the following parameter space

(1)

$$\mathcal{F}_\alpha = \left\{ \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le Mk^{-\alpha} \text{ for all } k, \text{ and } \lambda_{\max}(\Sigma) \le M_0 \right\}$$

where $\lambda_{\max}(\Sigma)$ is the maximum eigenvalue of the matrix Σ, and $\alpha > 0$, $M > 0$ and $M_0 > 0$. Note that the smallest eigenvalue of any covariance matrix in the parameter space F_α is allowed to be 0 which is more general than the assumption in equation (5) of Bickel and Levina (2008a). The parameter α in (1), which essentially specifies the rate of decay for the covariances σ_{ij} as they move away from the diagonal, can be viewed as an analog of the smoothness parameter in nonparametric function estimation problems. The optimal rate of convergence for

estimating Σ over the parameter space $\mathcal{F}_\alpha(M_0, M)$ critically depends on the value of α.

The distribution of the X_i's is assumed to be subgaussian in the sense that there is $\rho > 0$ such that

$$\mathbb{P}\{|\mathbf{v}'(\mathbf{X}_1 - \mathbb{E}\mathbf{X}_1)| > t\} \le e^{-t^2\rho/2} \text{ for all } t > 0 \text{ and } \|\mathbf{v}\|_2 = 1. \tag{2}$$

Let $\mathcal{P}_\alpha = \mathcal{P}_\alpha(M_0, M, \rho)$ denote the set of distributions of $\mathbf{X}_1$ that satisfy (1) and (2). Write $a_n \asymp b_n$ if there are positive constants c and C independent of n such that $c \le a_n/b_n \le C$. For a matrix A its operator norm is defined as $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2$. We assume that $p \le \exp(\gamma n)$ for some constant $\gamma > 0$.

We have the following optimal rate of convergence for estimating the covariance matrix under the operator norm.

Theorem 1. *The minimax risk of estimating the covariance matrix Σ over the class $\mathcal{P}_\alpha$ given in (1) satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}_\alpha} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \asymp \min\left\{n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n},\ \frac{p}{n}\right\}. \tag{3}$$

The following result gives the minimax rate of convergence for estimating the covariance matrix Σ under the Frobenius norm based on the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$.

Theorem 2. *The minimax risk under the Frobenius norm satisfies*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}_\alpha} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \asymp \min\left\{n^{-\frac{2\alpha+1}{2(\alpha+1)}},\ \frac{p}{n}\right\} \tag{4}$$

In addition, we discussed adaptive estimation, estimation under other matrix norms and sparse (inverse) covariance matrices estimation. For instance, the minimax risks of estimating the covariance matrix Σ under the matrix l_1 norm satisfies

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{P}_\alpha} \mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_1^2 \asymp \min\left\{n^{-\frac{\alpha}{\alpha+1}} + \left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}},\ \frac{p^2}{n}\right\} \tag{5}$$

References

[1] P. J. Bickel and E. Levina, *Regularized estimation of large covariance ma- trices*, Ann. Statist. **36** (2008a), 199-227.

[2] P. J. Bickel and E. Levina, *Covariance regularization by thresholding*, Ann. Statist. **32** (2008b), 2577-2604.

[3] T. T. Cai and H. H. Zhou, *Minimax Estimation of Large Covariance Matrices under l_1-Norm*, submitted.

[4] T. T. Cai, C.-H. Zhang and H. H. Zhou, *Optimal rates of convergence for covariance matrix estimation*, Ann. Statist., to appear.

Reporter: Vladimir Panov

Participants

Prof. Dr. Nathalie Akakpo
Laboratoire de Mathematique
Universite Paris Sud (Paris XI)
Batiment 425
F-91405 Orsay Cedex

Dr. Sylvain Arlot
D.I.
Ecole Normale Superieure
45, rue d'Ulm
F-75230 Paris Cedex 05

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

Dr. Denis Belomestny
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

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

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

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

Dr. Gilles Blanchard
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin

Prof. Dr. Natalia Bochkina
School of Mathematics
University of Edinburgh
James Clerk Maxwell Bldg.
King's Buildings, Mayfield Road
GB-Edinburgh EH9 3JZ

Dominique Bontemps
Universite de Paris-Sud
Laboratoire de Mathematiques
Batiment 430
F-91405 Orsay Cedex

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

Prof. Fabienne Comte
Laboratoire MAP 5
Universite Paris Descartes
45, rue des Saints-Peres
F-75270 Paris Cedex 06

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

Prof. Dr. Arnak Dalalyan
CERTIS - Ecole Nationale des
Ponts et Chaussees
6, Ave. Blaise Pascal
Cite Descartes Champs-s-Marne
F-77455 Marne-la-Vallee Cedex 2

Dr. Thorsten Dickhaus
Technical University of Berlin
Department Computer Science
Franklinstr. 28/29
10587 Berlin

Dr. Elmar Diederichs
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Prof. Dr. Noureddine El Karoui
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
USA

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

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

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

Prof. Dr. Yuri Golubev
Centre de Mathematiques et
d'Informatique
Universite de Provence
39, Rue Joliot-Curie
F-13453 Marseille Cedex 13

Le-Minh Ho
Mathematisches Institut
Humboldt-Universität Berlin
Burgstr. 26
10099 Berlin

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

Prof. Dr. Vladimir Koltchinskii
School of Mathematics
Georgia Institute of Technology
686 Cherry Street
Atlanta , GA 30332-0160
USA

Prof. Dr. Erwan Le Pennec
U. F. R. de Mathematiques
Case 7012
Universite Paris 7
2, Place Jussieu
F-75251 Paris Cedex 05

Prof. Dr. Oleg Lepski
Centre de Mathematiques et
d'Informatique
Universite de Provence
39, Rue Joliot-Curie
F-13453 Marseille Cedex 13

Karim Lounici
Dept. of Pure Mathematics and
Mathematical Statistics
University of Cambridge
Wilberforce Road
GB-Cambridge CB3 0WB

Zongming Ma
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA

Hilmar Mai
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

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

Prof. Dr. Pascal Massart
Laboratoire de Mathematique
Universite Paris Sud (Paris XI)
Batiment 425
F-91405 Orsay Cedex

Prof. Dr. Michael H. Neumann
Institut für Stochastik
Friedrich-Schiller-Universität
Ernst-Abbe-Platz 2
07743 Jena

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

Vladimir Panov
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

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

Prof. Dr. Patricia Reynaud-Bouret
Laboratoire de Mathematiques
Universite de Nice
Parc Valrose
F-06108 Nice Cedex

Prof. Dr. Vincent Rivoirard
Laboratoire de Mathematique
Universite Paris Sud (Paris XI)
Batiment 425
F-91405 Orsay Cedex

Dr. Angelika Rohde
Department Mathematik
Universität Hamburg
Bundesstr. 55
20146 Hamburg

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

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

Nora Serdyukova
Weierstraß-Institut für
Angewandte Analysis und Stochastik
im Forschungsverbund Berlin e.V.
Mohrenstr. 39
10117 Berlin

Jakob Söhl
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin

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

Claudia Strauch
Department Mathematik
Universität Hamburg
Bundesstr. 55
20146 Hamburg

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

Prof. Dr. Aad W. van der Vaart
Faculteit Wiskunde en Informatica
Vrije Universiteit Amsterdam
De Boelelaan 1081 a
NL-1081 HV Amsterdam

Prof. Dr. Martin Wainwright
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
USA

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

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