Mathematisches Forschungsinstitut Oberwolfach

Report No. 16/2009

DOI: 10.4171/OWR/2009/16

Sparse Recovery Problems in High Dimensions: Statistical Inference and Learning Theory

Organised by
Peter Bartlett, Berkeley
Vladimir Koltchinskii, Atlanta
Alexandre Tsybakov, Paris
Sara van der Geer, Zuerich

March 15th – March 21st, 2009

Abstract. The statistical analysis of high dimensional data requires new techniques, extending results from nonparametric statistics, analysis, probability, approximation theory, and theoretical computer science. The main problem is how to unveil, (or to mimic performance of) sparse models for the data. Sparsity is generally meant in terms of the number of variables included, but may also be described in terms of smoothness, entropy, or geometric structures. A key objective is to adapt to unknown sparsity, yet keeping computational feasibility.

Mathematics Subject Classification (2000): 62-06.

Introduction by the Organisers

In this workshop, experts from a wide range of mathematics shared their view on sparsity and presented an interesting blend of talks. The approaches discussed include exploiting a priori known structures, such as grouping of variables or graphical hierarchies, and the application of algorithms freed from the bodice of convexity. High dimensional problems lead to deep mathematical questions, and answers from often unexpected angles. The variety of perspectives that came up during this workshop made it into an truly inspiring experience.

Workshop: Sparse Recovery Problems in High Dimensions: Statistical Inference and Learning Theory

Table of Contents

Abstracts

Universal selection rule in non-parametric estimation and uniform bounds for norms of sums of independent random functions

Oleg Lepski
(joint work with A. Goldenshluger)

The talk consists of two parts.

Part I. This part is devoted to discussion of a new approach to nonparametric estimation which is based on selection from a given family of *linear estimators.* Our methodology is applied in various statistical settings since there is no restrictions related to the statistical model.

Let $\big(\mathcal{X}^{(n)}, \mathfrak{B}^{(n)}, \mathbb{P}_f^{(n)}, F \in \mathbb{F}\big)$ be a family of statistical experiments generated by an observation $X^{(n)}$. This means that $\mathfrak{B}^{(n)}$ is the σ-algebra generated by the random element $X^{(n)}$ and, the probability law of $X^{(n)}$ belongs to the family $\big(\mathbb{P}_f^{(n)}, f \in \mathbb{F}\big)$.

Let $\mathcal{D}$ be an open interval in $\mathbb{R}^d$, $d \geq 1$, let $\mathbb{F}$ be a set of Borel functions $f : \mathcal{D} \to \mathbb{R}$, and let $\mathfrak{m}$ be a σ-finite measure on $\mathcal{D}$. Our goal is to estimate the function f. To avoid discussion of boundary effects we are interested in estimating f on $\mathcal{D}_0$, where $\mathcal{D}_0$ is an open subinterval of $\mathcal{D}$. By an estimator of f we mean any $\mathcal{B}^{(n)}$-measurable mapping, $\hat{f} : \mathcal{X}^{(n)} \times \mathcal{D} \to \mathbb{F}_0$, where $\mathbb{F}_0 \supseteq \mathbb{F}$ is a separable linear metric space of functions defined on $\mathcal{D}$ and acting to $\mathbb{R}$. With any estimator we associate the risk

$$\mathcal{R}_n^{(\ell)}[\hat{f}; f] = \Big(\mathbb{E}_f^{(n)}\ \big[\ell(\hat{f} - f)\big]^q\Big)^{\frac{1}{q}},$$

where $\ell : \mathbb{F}_0 \to \mathbb{R}_+$ is a semi-norm, and $q > 0$ is a given real number. The problem is to estimate f from the observation $X^{(n)}$ with small risk $\mathcal{R}_\ell^{(n)}[\hat{f}; f]$, at least for large n.

We say that the estimator $\hat{f}(x)$ is linear *if there exists a function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that*

$$\mathbb{E}_f\big[\hat{f}(x)\big] = \int_{\mathcal{D}} K(t,x) f(t) \mathfrak{m}(\mathrm{d}t), \quad \forall F \in \mathbb{F},\ \forall x \in \mathcal{D}.$$

Thus, the linear estimator is the estimator whose expectation is a linear functional of the underlying function f. Let $\mathcal{D}_1$ be an open interval such that $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \mathcal{D}$.

Any function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\begin{aligned} \int_{\mathcal{D}} K(t,x)\mathfrak{m}(\mathrm{d}t) &= 1,\ \forall x \in \mathcal{D}_1; \\ supp\big(K(\cdot,x)\big) &\subseteq \mathcal{D}_1,\ \forall x \in \mathcal{D}_0, \end{aligned}$$

*will be called the $\mathcal{D}_1$-*weight*. Let $\mathfrak{K}(\mathcal{D}_1)$ be the set of all $\mathcal{D}_1$-weights.*

We endow $\mathfrak{K}(\mathcal{D}_1)$ with the operation "$\otimes$": $\forall K_1, K_2 \in \mathfrak{K}(\mathcal{D}_1)$

$$[K_1 \otimes K_2](\cdot,\cdot) = \int_{\mathcal{D}_1} K_1(\cdot, y)K_2(y,\cdot)\mathfrak{m}(\mathrm{d}y),$$

and we say that K_1 and K_2 commutate if

$$[K_1 \otimes K_2] \equiv [K_2 \otimes K_1].$$

A subset $\mathcal{K} \subset \mathfrak{K}(\mathcal{D}_1)$ *is called the* commutative weight system *if any pair of its elements commutate.*

Let $\mathcal{K}$ be a commutative weight system, and let $\mathcal{L}_{\mathcal{K}} = \mathcal{K}^{\otimes} \cup \mathcal{K}$, where

$$\mathcal{K}^{\otimes} = \left\{L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} :\ L = K \otimes K',\ K, K' \in \mathcal{K}\right\}.$$

Suppose that $\left(\mathcal{X}^{(n)}, \mathfrak{B}^{(n)}, \mathbb{P}^{(n)}_f, f \in \mathbb{F}\right)$ is the $\mathcal{L}_{\mathcal{K}}$-experiment, i.e a linear estimator is defined for any function belonging to $\mathcal{L}_{\mathcal{K}}$.

Let $\mathcal{F}_{\mathcal{K}} = \{\hat{f}_K,\ \ K \in \mathcal{K}\}$ be the collection of linear estimators generated by $\mathcal{K}$. We propose an estimator, say $f^\star$, for the underlying function f whose construction is based on the data-driven selection from the family $\mathcal{F}_{\mathcal{K}}$, namely

$$f^\star = \hat{f}_{\hat{K}},\ \ \hat{K} = \hat{K}_{X^{(n)}} \in \mathcal{K}.$$

We also establish an explicit upper bound on $\mathcal{R}^{(\ell)}_n[f^\star; f]$ for any given $f \in \mathbb{F}$ and $n \in \mathbb{N}^*$ in the case of an arbitrary semi-norm ℓ.

The remarkable property of our selection rule is that under rather mild technical assumptions it can be applied to any commutative weight system.

Construction of our selection rule requires finding uniform bounds on rather general stochastic processes. Their description and corresponding results are discussed in the second part of the talk.

Part II. In this part of the talk we present upper bounds for norms of random functions of special type.

Let $\left(\mathcal{T}, \mathfrak{T}, \tau\right)$ and $\left(\mathcal{X}, \mathfrak{X}, \varkappa\right)$ be σ-finite spaces, $\mathcal{X}$ be a Banach space, and let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a complete probability space.

Let X be a $\mathcal{X}$-valued random element defined on $(\Omega, \mathfrak{A}, \mathbb{P})$ and having the density f with respect to the measure $\varkappa$. Let also ε be a real random variable defined on the same probability space which is independent of X and has symmetric distribution.

For any $\left(\mathfrak{T} \times \mathfrak{X}\right)$-measurable function w on $\mathcal{T} \times \mathcal{X}$ and for any $t \in \mathcal{T},\ \ n \in \mathbb{N}^*$ we define

$$\xi_w(t) = \sum_{i=1}^{n} \left[w\big(t, X_i\big) - \mathbb{E}w(t, X)\right], \quad \eta_w(t) = \sum_{i=1}^{n} w(t, X_i)\varepsilon_i,$$

where $(X_i, \varepsilon_i), i = \overline{1,n}$, are *independent* copies of (X, ε).

Put for $1 \le s < \infty$

$$\left\|\xi_w\right\|_{s,\tau} = \left[\int \left|\xi_w(t)\right|^s \tau(\mathrm{d}t)\right]^{\frac{1}{s}}, \quad \left\|\eta_w\right\|_{s,\tau} = \left[\int \left|\eta_w(t)\right|^s \tau(\mathrm{d}t)\right]^{\frac{1}{s}},$$

and let W be a given set of $(\mathfrak{T} \times \mathfrak{X})$-measurable functions.

Let Ψ_w be either ξ_w or η_w. We are interested in finding a *non-random* function on W, say $U_\Psi(w)$, which is the *uniform* upper bound on $\big\|\Psi_w\big\|_{s,\tau}$ in the sense that

$$\mathbb{P}\left\{\sup_{w\in W}\Big[\big\|\Psi_w\big\|_{s,\tau} - \mathsf{u}C^*(y)U_\Psi(w)\Big] \geq 0\right\}$$

is small and tends to zero as $n \to \infty$ for any fixed $y > 0$. Here $C^*(\cdot)$ is the *given* linear function, and $\mathsf{u} \geq 1$ is the constant that is completely determined by W and often $\mathsf{u} = 1$.

In fact we want to bound from above the latter probability as well as the expectation

$$\mathbb{E}\left(\sup_{w\in W}\Big[\big\|\Psi_w\big\|_{s,\tau} - \mathsf{u}C^*(y)U_\Psi(w)\Big]\right)_+^q, \ q \geq 1.$$

We provide explicit expressions for U_Ψ which are completely determined by w, f and s. We also show that in the case of $\Psi_w = \eta_w$ the corresponding uniform bound depends on the moment conditions on the distribution of ε.

Another problem arising in applications of the obtained results in mathematical statistics is to find a uniform bound independent of the density f. Sometimes the dependence on f is not crucial because the function f is supposed to be known. The typical example is the regression model where the random function η_w appears.

There exist, however, problems where the situation is completely different. One of them is the problem of estimating a unknown multivariate density from *i.i.d.* observations, where the process ξ_w appears. In this case ξ_w can be treated as the stochastic error of the linear estimator associated with the weight function w. A uniform non-random bound is used in the selection rule, discussed in Part I; it allows to select the estimator from a given family of linear estimators. It is clear that the use of a uniform bound depending on the unknown parameter is impossible for this purpose. To overcome this difficulty we propose a random uniform bound, say $\hat{U}_s(w)$, whose construction is based only on the sequence $X_1, \dots, X_n$, and establish corresponding inequality for

$$\mathbb{E}\left(\sup_{w\in W}\Big[\Psi_w - 2\mathsf{u}C^*(y)\hat{U}_s(w)\Big]\right)_+^q, \ q > 0.$$

The obtained result, together with approach developed in Part I is sufficient in order to establish a general *oracle inequality* in the context of multivariate density estimation. In particular, it allows to solve completely the problem of *bandwidth selection* for the risks described by $\mathbb{L}_s$-norm. The solution was known only for $s = 1$ and it was obtained by means of absolutely different technique. It allows also to construct an estimator which is *adaptive* with respect to the anisotropic Sobolev classes estimator. This problem was solved only for $s = 2$ and $s = \infty$.

Stability Selection

Nicolai Meinshausen
(joint work with Peter Bühlmann)

Estimation of structure, such as in graphical modeling, cluster analysis or variable selection, is notoriously difficult, especially for high-dimensional data. We introduce the new method of stability selection. It is based on subsampling in combination with (high-dimensional) selection algorithms. As such, the method is extremely general and has a very wide range of applicability. Stability selection provides finite sample control for some error rates of false discoveries and hence a transparent principle to choose a proper amount of regularisation for structure estimation or model selection. Maybe even more importantly, results are typically remarkably insensitive to the chosen amount of regularisation. Another property of stability selection is the improvement over a pre-specified selection method. We prove for randomized Lasso that stability selection will be variable selection consistent even if the necessary conditions needed for consistency of the original Lasso method are violated. We demonstrate stability selection for variable selection and Gaussian graphical modeling, using motif regression data and some simulated examples.

High-dimensional intervention effects and causality

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

Our work is motivated by the following problem in biology. We want to know which genes play a role in a certain phenotype, say a disease status or, in one of our cases, a continuous value of riboflavin (vitamin B2) production in the bacterium Bacillus Subtilis. To be more precise, our goal is to infer which genes have an effect on the phenotype in terms of an intervention: if we knocked down single genes, which of them would show a relevant or important effect on the phenotype? The difficulty is, however, that the available data are only observational. Using such observational data, we want to infer all (single gene) intervention effects. This task coincides with inferring causal effects, a well-established area in statistics, cf. [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 demonstrate the method to predict the most important intervention effects in two large-scale biological systems from Bacillus Subtilis and S.Cerevisiae.

References

[1] J. Pearl, *Causality: models, reasoning and inference.* Cambridge University Press, 2000.

[2] P. Spirtes, 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, M. Kalisch and P. Bühlmann, *Estimating high-dimensional intervention effects from observational data.* The Annals of Statistics, to appear.

On verifiable conditions of ℓ_1-recovery of sparse signals with sign restrictions

Anatoli Juditsky

(joint work with F. Kilinc Karzan and A. Nemirovski)

We address the recovery problem as follows: assess a *sparse* signal $w \in \mathbb{R}^n$ with *sign restrictions* given an observation $y \in \mathbb{R}^m$:

$$y = Aw + e, \quad \|e\| \leq \varepsilon,$$

where $A \in \mathbb{R}^{m \times n}$ (in this context $m < n$) is a given matrix, $\|\cdot\|$ is a given norm on $\mathbb{R}^m$, e is the observation error and $\varepsilon \geq 0$ is a given upper bound on the error magnitude, measured in the norm $\|\cdot\|$. A popular solution to the problem is given by the ℓ_1-recovery, which amounts to take as estimation of w as an optimal solution $\widehat{w}$ to the optimization problem

$$(1)\quad \widehat{w} \in \operatorname*{arg\,min}_x \{\|x\|_1 : \ \|Ax - y\| \leq \varepsilon, \ x_i \geq 0 \ \forall i \in P_+, \ x_i \leq 0 \ \forall i \in P_-\}$$

(here P_+, P_- are the subsets of $\{1, ..., n\}$ and $P_+ \cap P_- = \emptyset$). Note that when $P_+ = P_- = \emptyset$, this problem reduces to the most commonly studied estimator in the existing *Compressive Sensing* theory.

Our objective is given a $m \times n$ matrix A to answer (efficiently) the question if the matrix A is such that whenever the true signal w is s-sparse, the ℓ_1-recovery

$$\widehat{w}' \in \operatorname*{arg\,min}_x \{\|x\|_1 : \ Ax = Aw, \ x_i \geq 0 \ \forall i \in P_+, \ x_i \leq 0 \ \forall i \in P_-\}$$

recovers w exactly? In the case when the answer is positive, we say that A is s-*semigood.*

We develop several equivalent necessary and sufficient conditions of s-semigoodness of the sensing matrix A in the spirit of [2]. Then we provide the bounds for the accuracy of inexact ℓ_1-recovery (case of noisy observation and optimization problem (1) solved up to approximate optimality) in terms of the quantities which participate in these conditions.

We use the LP-relaxation technique, introduced in [1], to design new *verifiable* sufficient conditions for s-semigoodness of A and provide an upper bound on maximal s such that A is s-semigood. We provide as well a sufficient condition for s-semigoodness of A based on Semidefinite Relaxation.

The verifiable sufficient condition in question allows us to establish a direct link between the ℓ_1-recovery and the linear recovery and develop a new Matching Pursuit algorithm for iterative improvement of linear recovery for sparse signals.

References

[1] A. Juditsky and A. Nemirovski, *On Verifiable Sufficient Conditions for Sparse Signal Recovery via ℓ_1 Minimization*, `http://hal.archives-ouvertes.fr/hal-00321775/` . Submitted to *Mathematical Programming* (2008).

[2] Y. Zhang, *A simple proof for recoverability of ℓ_1-minimization (II): the non-negative case.* Technical report TR05-10, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2005.

LOL : Learning Optimal Leaders

Dominique Picard

(joint work with Gérard Kerkyacharian, Mathide Mougeot and Karine Tribouley)

In this paper, we are interested in the problem of learning an unknown real valued function defined a compact domain in $\mathbb{R}^d$.

One of our purposes will be to link this problem to a general approach on high dimensional linear models in statistics and propose some tools resulting from a combination of inspirations.

We assume to observe an n sample $Z_1, \dots, Z_n$ of $Z = (X, Y)$. The distribution of Z is denoted by ρ. Our aim is to recover the function f:

$$f(x) = E_\rho[Y|X = x].$$

We describe a procedure which is

- Very Simple to implement. (no Argmin...)
- Universal (adaptive)
- Has optimal exponential error bounds

Description of the LOL procedure:
We consider a (finite) dictionary of size p (can also come from a kernel...)

$$\mathcal{D} = \{g\} \subset \mathbb{L}_2(\hat{\rho})$$

which we normalize : $\|g\|_{\hat{\rho}} = 1$, for all g.

- we calculate the coherence of the dictionary

$$\tau_n := \sup_{g,g' \in \mathcal{D},\ g \neq g'} \frac{1}{n} \sum_{i=1}^{n} g(X_i) g'(X_i)$$

- this provides us with a quantity : $N = [\delta/\tau_n]$. ($0 < \delta < 1$ is a fixed real number associated to the procedure -for instance $\delta = 1/2$-)

Notice that each time we consider a set of $m \leq N$ different vectors of the dictionary

$$\{g_1, \dots g_m\}$$

the $m \times m$ matrix $\frac{1}{n} G_m G_m^t$ with entries, $[\frac{1}{n}\sum_{i=1}^n g_k(X_i) g_l(X_i)]_{kl,\ 1 \leq k,l \leq m}$ is almost diagonal, in the sense that : (Restricted Isometry Property)

$$\forall x \in \mathbb{R}^m,\ \ \|x\|_{l_2}^2 (1-\delta) \leq x^t [\frac{1}{n} G_m G_m^t] x \leq \|x\|_{l_2}^2 (1+\delta) \tag{1}$$

- Fix $\lambda^1 = T_1 (\frac{\log p}{n})^{\frac{1}{2}}$.
- Find the set $A = \{g \in \mathcal{D},\ |\frac{1}{n}\sum_{i=1}^n g(X_i) Y_i| \geq \lambda^1\}$
- If $Card(A) \leq N$ we keep the whole set $B = A$
- If $Card(A) \geq N$ we take the N largest $B \subset A$
- Consider the pseudo-regression model :

$$Y_i = \sum_{g \in B} \alpha_g g(X_i) + \epsilon_i$$

• Let $\hat{\alpha} = (\hat{\alpha}_g,\ g \in B)$ be the minimum least square error in this model :

$$\sum_{i=1}^{n} (Y_i - \sum_{g \in B} \hat{\alpha}_g g(X_i))^2 \text{ minimum}$$

$$\hat{\alpha} = (G_B G_B^t)^{-1} G_B Y$$

with $Y = (Y_1, \dots, Y_n)^t$
$(G_B)_{li} = g_l(X_i),\ l \in B,\ i \in \ \{1, \dots, n\}.$

• We obtain our estimator using a final thresholding:

$$\hat{f}(x) := \sum_{g \in B} \hat{\alpha}_g g I\{|\hat{\alpha}_g| \geq T_2 \sqrt{\frac{\log n}{n}}\}$$

Theorem 1. *In Gaussian regression framework, if :*
$\tau_n \leq c\sqrt{\frac{\log n}{n}}$
$Card(\mathcal{D}) \leq n^a,\ T_1 \geq c_1(\delta, M),\ T_2 \geq c_2(\delta)$
There exist positive constants D and γ, such that

$$\sup_{\rho,\ f \in V} \rho^{\otimes}\{\|f - \hat{f}\|_{\hat{\rho}} > \eta\} \leq \{ \begin{array}{ll} e^{-\gamma n \eta^2}, & \eta \geq D\eta_n, \\ 1, & \eta \leq D\eta_n \end{array} \tag{2}$$

$$\eta_n^2 = [\frac{S}{n}].$$

V is defined by sparsity conditions on f, depending on S, M.

Classification of sparse high-dimensional vectors

Christophe Florent Pouet
(joint work with Yuri I. Ingster and Alexandre B. Tsybakov)

Let $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ be two i.i.d. samples from two different populations with probability distributions P_X and P_Y on $\mathbb{R}^d$ respectively. Here $X_i = (X_i^1, \dots, X_i^d)$, $\quad Y_j = (Y_j^1, \dots, Y_j^d)$ where X_i^k and Y_j^k are the components of X_i and Y_j. We assume that

$$X_i^k = \xi_i^k, \qquad Y_j^k = u_k + \eta_j^k, \tag{1}$$

where $u = (u_1, \dots, u_d)$ is a deterministic mean vector and the errors $\xi_i^1, \dots, \xi_i^d$, $\eta_j^1, \dots, \eta_j^d$ are jointly i.i.d. zero mean random variables with probability density f on $\mathbb{R}$.

We consider the problem of discriminant analysis when the dimension of the observations d is very large (tends to $+\infty$). Assume that we observe a random vector $Z = (Z^1, \dots, Z^d)$ independent of $(\mathbf{X}, \mathbf{Y})$ and we know that the distribution of Z is either P_X or P_Y. Our aim is to classify Z, i.e., to decide whether Z comes from the population with distribution P_X or from that with distribution P_Y. A full review of our results is available in [4].

As in [5, 1], we introduce a set of sparse vectors in $\mathbb{R}^d$ characterized by a number $a_d > 0$ and a *sparsity index* $\beta \in (0,1]$ to measure the closeness between u and 0 :

$$U_{\beta,a_d} = \left\{ u = (u_1, \dots, u_d) : \ u_k = a_d \eta_k, \ \eta_k \in \{0,1\}, \ cd^{1-\beta} \leq \sum_{k=1}^{d} \eta_k \leq Cd^{1-\beta} \right\},$$

where $0 < c < C < +\infty$ are two constants.

Let $\psi = \psi(\mathbf{X}, \mathbf{Y}, Z) \in [0, \ 1]$ be a decision rule. If $\psi = 0$ we allocate Z to the P_X-population, whereas for $\psi = 1$ we allocate Z to the P_Y-population. If f is known, we do not need the sample $\mathbf{X}$ to construct decision rules. Let $P_{H_0}^{(u)}$ and $P_{H_1}^{(u)}$ denote the joint probability distributions of $\mathbf{X}, \mathbf{Y}, Z$ when $Z \sim P_X$ and $Z \sim P_Y$ respectively, and let $E_{H_0}^{(u)}$ and $E_{H_1}^{(u)}$ denote the corresponding expectations. Consider the maximum risk

$$\mathcal{R}_M(\psi) = \max\Big(E_{H_0}^{(u)}(\psi), E_{H_1}^{(u)}(1-\psi) \Big).$$

The *classification boundary* is the condition on (β, a_d) corresponding to the passage from the fact that *successful classification is possible*, i.e., β and a_d are such that

$$\lim_{d\to+\infty} \inf_{\psi} \sup_{u \in U_{\beta,a_d}} \mathcal{R}_M(\psi) = 0,$$

to the fact that *successful classification is impossible*,i.e., β and a_d are such that

$$\liminf_{d\to+\infty} \inf_{\psi} \sup_{u \in U_{\beta,a_d}} \mathcal{R}_M(\psi) = 1/2.$$

We assume that there exists $0 \leq \gamma < 1$ such that $\lim_{d\to\infty} (\log m)/(\log d) = \gamma$. According to the value of β, we distinguish between *moderately sparse vectors* $(0 < \beta \leq (1-\gamma)/2)$ and *highly sparse vectors* $((1-\gamma)/2 < \beta < 1-\gamma)$.

The classification boundary for moderately sparse vectors is of the form

$$R_d \triangleq d^{1/2-\beta} a_d \asymp 1. \tag{2}$$

Moreover, (2) holds under weak assumptions on the density f of the noise.

In the case of highly sparse vectors, we consider Gaussian or approximately Gaussian noise distributions which leads us to a special dependence of a_d on d and m of the form $a_d \asymp \sqrt{(\log d)/m}$.

When $m \geq 1$ is a fixed integer, and the noise density f is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$, the set of highly sparse vectors corresponds to $\beta \in (1/2, 1)$. We consider $a_d = s\sigma\sqrt{\log d}$, where $s > 0$ is fixed. The classification boundary is expressed by the following condition on β, s and m :

$$\begin{aligned} s\sqrt{m+1} &= \phi(\beta), \\ \text{where} \quad \phi(\beta) &= \begin{cases} \sqrt{2\beta - 1} & \text{if } 1/2 < \beta \leq 3/4, \\ \sqrt{2}\Big(1 - \sqrt{1-\beta}\Big) & \text{if } 3/4 < \beta < 1. \end{cases} \end{aligned}$$

This classification boundary is also extended to the case where s depends on d so that $s\sqrt{m+1}$ stays bounded.

When $m \to +\infty$ as $d \to +\infty$, $(\log m)/(\log d) \to \gamma \in [0,1)$ and f is Gaussian $\mathcal{N}(0,\sigma^2)$ with known or unknown $\sigma > 0$ we consider $a_d = x\sigma\sqrt{(\log d)/m}$, where $x > 0$ is fixed. Then for $\beta > 1-\gamma$ successful classification is impossible. For $(1-\gamma)/2 < \beta < 1-\gamma$ (the highly sparse zone), the classification boundary is of the form

$$x/\sqrt{1-\gamma} = \phi(\beta/(1-\gamma)).$$

The upper bounds are extended to the case where $m \to +\infty$ as $d \to +\infty$, $(\log m)/(\log d) \to \gamma \in [0,1)$, $m/\log d \to +\infty$, and the noise satisfies the Cramér condition.

In a work parallel to ours, Donoho and Jin [2, 3] and Jin [6] independently have analysed a setting less general than the present one, without considering a minimax framework. They demonstrated that the higher criticism (HC) methodology can be successfully extended to the classification problem.

References

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

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

[3] D. Donoho and J. Jin, *Feature selection by higher criticism thresholding: Optimal phase diagram.* Manuscript, available at arXiv:0812.2263.

[4] Y. Ingster, C. Pouet and A.B. Tsybakov, *Sparse classification boundaries*, Manuscript, avalaible at arXiv:0903.4807.

[5] Y. Ingster and I. Suslina, *Nonparametric Goodness-of-Fit Testing under Gaussian Model.* Springer Lectures Notes in Statistics. Vol. **169** (2002), Springer, New York.

[6] J. Jin, *Impossibility of successful classification when useful features are rare and weak.* Manuscript (2009).

Largest eigenvalues and eigenvectors in multivariate analysis

Iain Johnstone

The talk has two parts. We first review the role of sample eigenvalues in some classical methods of multivariate statistics, such as principal components and canonical correlation analysis. Results from random matrix theory can provide practically useful approximations when the ratio of the number of variables (p and q) to sample size n is not necessarily small. For concreteness, we focus on the limiting distribution for the largest principal component variance and the largest canonical correlation in "null hypothesis" settings when the data matrices have independent standard Gaussian entries, though we also give a rate of convergence result for the largest eigenvalue of square symmetric Gaussian matrices. We also review concentration inequalities for the largest eigenvalue of a white Wishart matrix and propose a concentration bound for the "double Wishart" setting exemplified by canonical correlation analysis.

In the second part of the talk we turn to estimation of the leading eigenvectors in a "spiked" covariance model of the form $\Sigma = \sigma^2 I + \sum_{\nu=1}^{M} \lambda_\nu \theta_\nu \theta_\nu^T$. This work is joint with Debashis Paul, University of California at Davis [4]. The observations are assumed to be independent Gaussian random vectors, and the dimension increases to infinity as the sample size increases. We establish a lower bound on the rate of convergence of the minimax risk of the estimators under an L^2 loss on the eigenvectors that describes three different regimes of sparsity. We propose a new method for estimating the eigenvectors that is based on a two-stage coordinate selection scheme, assuming a certain degree of sparsity. We prove that under appropriate regularity conditions, the proposed estimator attains the optimal rate of convergence. We demonstrate the practical performance with a simulation study.

Following is a more detailed statement of one result to be discussed in the eigenvalue section of the talk. Further details may be found in Johnstone [3] and an applications oriented paper Johnstone [2]. Let $A \sim W_p(I, m)$ follow a Wishart distribution, independent of $B \sim W_p(I, n)$, where $m \geq p$. Then the largest eigenvalue θ of $(A+B)^{-1}B$ is called the *greatest root statistic* and a random variate having this distribution is denoted $\theta_1(p, m, n)$, or $\theta_{1,p}$ for short. Equivalently $\theta_1(p, m, n)$ is the largest root of the determinantal equation

$$\det[B - \theta(A + B)] = 0. \tag{1}$$

In general the parameter p refers to dimension, m to the "error" degrees of freedom and n to the "hypothesis" degrees of freedom. Thus $m + n$ represents the "total" degrees of freedom.

Assume p is even and that $p, m = m(p)$ and $n = n(p) \to \infty$ together in such a way that

$$\lim_{p\to\infty} \frac{\min(p, n)}{m + n} > 0, \qquad \lim_{p\to\infty} \frac{p}{m} < 1. \tag{2}$$

A consequence of our main result, stated more completely below, is that with appropriate centering and scaling, the logit transform $W_p = \text{logit}\, \theta_{1,p} = \log(\theta_{1,p}/(1-\theta_{1,p}))$ is approximately Tracy-Widom distributed:

$$\frac{W_p - \mu_p}{\sigma_p} \quad \overset{\mathcal{D}}{\Rightarrow} \quad Z_1 \sim F_1. \tag{3}$$

The distribution F_1 was found by [5] as the limiting law of the largest eigenvalue of a p by p Gaussian symmetric matrix; further information on F_1 is reviewed, for example, in [1].

The centering and scaling parameters are given by

$$\begin{aligned}\mu_p &= 2\log\tan\left(\frac{\varphi+\gamma}{2}\right),\\ \sigma_p^3 &= \frac{16}{(m+n-1)^2}\,\frac{1}{\sin^2(\varphi+\gamma)\sin\varphi\sin\gamma},\end{aligned} \tag{4}$$

where the angle parameters γ, φ are defined by

$$\begin{aligned}\sin^2\left(\frac{\gamma}{2}\right) &= \frac{\min(p,n)-\frac{1}{2}}{m+n-1},\\ \sin^2\left(\frac{\varphi}{2}\right) &= \frac{\max(p,n)-\frac{1}{2}}{m+n-1}.\end{aligned} \tag{5}$$

Theorem 2. *Assume that $m(p), n(p) \to \infty$ as $p \to \infty$ through even values of p according to (2). For each $s_0 \in \mathbb{R}$, there exists $C > 0$ such that for $s \geq s_0$,*

$$|P\{W_p \leq \mu_p + \sigma_p s\} - F_1(s)| \leq Cp^{-2/3}e^{-s/2}.$$

Here C depends on (γ, φ) and also on s_0 if $s_0 < 0$.

The "correction factors" of $-\frac{1}{2}$ and -1 yield a second order rate of convergence that has important consequences for the utility of the approximation in practice. Indeed, we argue that it is reasonably accurate over the entire range of (non-asymptotic) values of the parameters. "Reasonably accurate" means, for example, less than ten percent relative error in the 95th percentile, even when working with two variables and any combination of error and hypothesis degrees of freedom.

We present numerical evidence that, for many applied purposes, the Tracy-Widom approximation presented here can often, if not quite always, substitute for the elaborate tables and computational procedures that have until now been needed.

References

[1] Johnstone, I. M. (2001), *On the distribution of the largest eigenvalue in principal components analysis.* Annals of Statistics **29**, 295–327.

[2] Johnstone, I. M. (2008*a*), *Approximate null distribution of the largest root in multivariate analysis.* Annals of Applied Statistics. To appear.

[3] Johnstone, I. M. (2008*b*), *Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence.* Annals of Statistics **36**(6), 2638–2716.

[4] Paul, D. and Johnstone, I. (2007), *Sparse principal component analysis for high dimensional data.* Technical report, UC Davis. Manuscript in progress.

[5] Tracy, C. A. and Widom, H. (1996), *On orthogonal and symplectic matrix ensembles.* Communications in Mathematical Physics **177**, 727–754.

Sparse inverse covariance estimation by minimizing L1-penalized log-determinant divergence

Bin Yu

(joint work with Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, Pradeep Ravikumar, Vince Vu, Yuval Benjamini, Kendrick Kay, Thomas Naslaries and Jack Gallant)

(The first part of the talk is based on joint work with Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti in statistics at UC Berkeley; the second part of the talk is based on joint work with Pradeep Ravikumar, Vince Vu, Yuval Benjamini in statistics at UC Berkeley and Kendrick Kay, Thomas Naslaries and Jack Gallant from the Helen Wills Neuroscience Institute at UC Berkeley.)

Extracting useful information from high-dimensional data is the focus of today's statistical research and practice. After broad success of statistical machine learning on prediction through regularization, interpretability is gaining attention and sparsity has been used as its proxy. With the virtues of both regularization and sparsity, L1 penalized minimization has been very popular recently.

In this talk, I would like to discuss the theory and pratcice of sparse inverse covariance modeling. The first part is on sparse inverse covariance estimation based on L1 penalized negative Gaussian log likelihood (or for the general case, the log determinant Bregman divergence). Given i.i.d. observations of a random vector in p-dim, we study the problem of estimating both its covariance matrix and its inverse covariance or concentration matrix. We estimate the inverse cov. matrix by minimizing an L1-penalized log-determinant Bregman divergence; in the multivariate Gaussian case, this approach corresponds to L1-penalized maximum likelihood, and the structure of inverse cov. is specified by the graph of an associated Gaussian Markov random field. We analyze the performance of this estimator under high-dimensional scaling, in which the number of nodes in the graph p, the number of edges s and the maximum node degree d, are allowed to grow as a function of the sample size n. In addition to the parameters (p,s,d), our analysis identifies other key parameters values in the true model that control rates. Our first result establishes consistency of our estimate for the inverse cov. in the elementwise maximum-norm. This in turn allows us to derive convergence rates in Frobenius and spectral norms, with improvements upon existing results for graphs with maximum node degrees. In our second result, we show that with probability converging to one, the estimated inverse cov. correctly specifies the zero pattern of the concentration matrix. We illustrate our theoretical results via simulations for various graphs and problem parameters, showing good correspondences between the theoretical predictions and behavior in simulations.

The second part is on collaborative research with the Gallant Lab at Berkeley on building sparse models that describe fMRI responses in primary visual cortex area V1 to natural images. The goal of the Gallant Lab is to undertand primate visual pathway. fMRI is an indirect measurement of neural activities and data have been collected by the Gallant Lab through stimulation of the visual pathway of a human subject with iid samaples of natural images from a database of over 11,000 images. Based on accepted biological understanding of primary visual cortex area V1, Gabor wavelet transforms of 128 by 128 circularly cropped natural images are obtained and fMRI signals over 1200 voxels in V1 responding to the images are gathered. When we started analyzing these data, the state-of-the-art method at the Gallant Lab was epsilon-L2boosting which is closedly related to Lasso (L1 penalized L2 minimization) based on a linear model with predictors as the Gabor transformed images with dim over p=10,000. We have n=1,750 images used as stimuli or we have 1,750 samples. The response variable in the linear model is pre-processed fMRI signal and each voxel is fitted a different model. Based on the nonlinear spase model SpAM of Ravikumar et al (2007), we discovered nonliner compressive properties of the fMRI signal and this nonlinear model lead to 12% prediction improvement on a separate 120 sample validation set (measured by correlation) over 1200+ voxels. A further modeling of constraining all the nonlinear transforms to be the same for each voxel lead to the new iV-SpAM model which brought a further 5% improvement. This high-power machine learning approach to discover the nonlinear property inspired later parametric power transforms (square root or 1/4 root) of the predictors and the prediciton performance improved over e-L2boosting by 21% with the parametrically transformed predictors. A preliminary modeling using sparse graphical model based on the fixed power transforms indicate promising directions to model voxels jointly. These models will be further validated through decoding of images using fMRI signals.

Inference in high-dimensional settings: Trade-offs between computational and statistical efficiency

Martin Wainwright

High-dimensional sparse PCA: Computational versus statistical efficiency. Principal component analysis (PCA) is a classical method for dimensionality reduction based on extracting the dominant eigenvectors of the sample covariance matrix (e.g., [7, 2]). However, as shown by Johnstone and collaborators [5, 6], standard PCA is well known to behave poorly in the "large p, small n" setting, in which the problem dimension p is comparable to or larger than the sample size n. In the paper [1], we study PCA in this high-dimensional regime, but under the additional assumption that the maximal eigenvector is sparse, say with at most k non-zero components. We consider the spiked covariance model [5] in which a base matrix is perturbed by adding a k-sparse maximal eigenvector, and analyze two computationally tractable methods for recovering the support set of this maximal eigenvector: (a) a simple thresholding method, originally used as a pre-processing step by Johnstone and Lu [6], which transitions from success to failure as a function of the rescaled sample size $\theta_{\mathrm{dia}}(n,p,k) = n/[k^2 \log(p-k)]$; and (b) a more sophisticated semidefinite programming (SDP) relaxation due to d'Asprémont et al. [3], which succeeds once the rescaled sample size $\theta_{\mathrm{sdp}}(n,p,k) = n/[k \log(p-k)]$ is larger than a critical threshold. In addition, we prove that no method, including the best method which has exponential-time complexity, can succeed in recovering the support if the order parameter $\theta_{\text{sdp}}(n, p, k)$ is below a threshold.

These results are complementary to the work of Paul and Johnstone [11], which focuses on recovery of sparse eigenvectors in ℓ_2 norm, as opposed to model selection consistency. Our results also highlight some interesting trade-offs between computational and statistical costs in high-dimensional inference. On one hand, the statistical efficiency of SDP relaxation is substantially greater than the thresholding method, requiring $\mathcal{O}(1/k)$ fewer observations to succeed. However, the computational complexity of SDP is also larger by roughly a factor $\mathcal{O}(p^3)$: one implementation due to d'Asprémont et al. has complexity $\mathcal{O}(np^2 + p^4 \log p)$ as opposed to the $\mathcal{O}(np^2 + p \log p)$ complexity of the thresholding method. Moreover, our information-theoretic analysis shows that the best possible method—namely, one based on an exhaustive search over all $\binom{p}{k}$ subsets, with exponential complexity—does not have substantially greater statistical efficiency than the SDP relaxation.

Structured regularization in multivariate regression: Benefits and dangers. In multivariate regression, a r-dimensional response vector is regressed upon a common set of p covariates, with a matrix $B^* \in \mathbb{R}^{p \times r}$ of regression coefficients. In various applications, it is natural to expect that the sparsity pattern matrix of the regression matrix B^* is block-structured, which suggests the use of block-structured regularization, as in a line of past work by various authors (e.g., [14, 13, 16, 15]). In contrast to the simple quadratic programs that arise with the Lasso [12], these block-structured regularizers typically lead to more general conic programs, including second-order cone programs (SOCPs) and semidefinite programs (SDPs). In this realm, some interesting questions include:

(a) when does the use of structured regularization (with its) higher computational cost) yield improved statistical efficiency?

(b) conversely, are their settings in which structured regularization and conic programming can impair statistical efficiency relative to computationally cheaper approaches?

In a line of recent work [10, 9], we have obtained some precise answers to these questions in certain settings. In addition, some other participants of the Oberwolfach workshop have recently obtained some related results [8, 4], with particular focus on the ℓ_1/ℓ_2 case.

In the paper [10], we study the behavior of the block-regularized ℓ_1/ℓ_2 Lasso for the problem of union support recovery, meaning recovery of the set of k rows for which B^* is non-zero. Under high-dimensional scaling, we show that the group Lasso recovers the exact row pattern with high probability over the random design and noise for (n, p, k) such that the *rescaled sample size* given by $\theta(n, p, k) := n/[2\psi(B^*) \log(p - k)]$ exceeds a critical threshold depending on the signal-to-noise ratio. Here n is the sample size, and $\psi(B^*)$ is a *sparsity-overlap function* measuring a combination of the sparsities and overlaps of the r-regression coefficient vectors that constitute the model. This sparsity-overlap function reveals that, if the design is uncorrelated on the active rows, ℓ_1/ℓ_2 regularization for multivariate regression never harms performance relative to an ordinary Lasso approach, and can yield substantial improvements in sample complexity (up to a factor of r) when the regression vectors are suitably orthogonal. For more general designs, it is possible for the ordinary Lasso to outperform the group Lasso, but these problems are not typical.

In the paper [9], we analyze the high-dimensional scaling of ℓ_1/ℓ_∞-regularized quadratic programming, considering both consistency rates in ℓ_∞-norm, and also how the minimal sample size n required for performing variable selection grows as a function of the model dimension, sparsity, and overlap between the supports. We begin by establishing bounds on the ℓ_∞-error as well sufficient conditions for exact variable selection for fixed design matrices, as well as designs drawn randomly from general Gaussian matrices, showing that high-dimensional scaling of ℓ_1/ℓ_∞-regularization is qualitatively similar to that of ordinary ℓ_1-regularization. Our second set of results applies to $r = 2$ linear regresion problems whose supports overlap in a fraction $\alpha \in [0, 1]$ of their entries and with design matrices drawn from standard Gaussian ensembles: for this problem class, we prove that the ℓ_1/ℓ_∞-regularized method undergoes a phase transition—that is, a sharp change from failure to success—characterized by the rescaled sample size $\theta_{1,\infty}(n, p, s, \alpha) = n/\{(4-3\alpha)s \log(p - (2 - \alpha)\, s)\}$. More precisely, given sequences of problems specified by (n, p, s, α), for any $\delta > 0$, the probability of successfully recovering both supports converges to 1 if $\theta_{1,\infty}(n, p, s, \alpha) > 1+\delta$, and converges to 0 for problem sequences for which $\theta_{1,\infty}(n, p, s, \alpha) < 1 - \delta$. An implication of this threshold is that use of ℓ_1/ℓ_∞-regularization yields improved statistical efficiency if the overlap parameter is large enough ($\alpha > 2/3$), but has *worse statistical efficiency* than a computationally less expensive Lasso-based approach for moderate to small overlap ($\alpha < 2/3$), or if the regression vectors are not close in absolute value on their common support.

References

[1] A. A. Amini and M. J. Wainwright, *High-dimensional analysis of semdefinite relaxations for sparse principal component analysis.* Annals of Statistics, to appear. Appeared as Tech. Report 747, Department of Statistics, UC Berkeley.

[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1984.

[3] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming.* SIAM Review, 49(3):434–448, July 2007.

[4] J. Huang and T. Zhang, *The benefit of group sparsity.* Technical Report arXiv:0901.2962, Rutgers University, January 2009.

[5] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis.* Annals of Statistics, 29(2):295–327, April 2001.

[6] I. M. Johnstone and A. Lu, *Sparse principal components.* Technical report, Stanford University, July 2004.

[7] I. T. Jolliffe, *Principal Component Analysis.* Springer, New York, NY, 2004.

[8] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer, *Taking advantage of sparsity in multi-task learning.* Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[9] S. Negahban and M. J. Wainwright, *Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization.* Technical report, Department of Statistics, UC Berkeley, March 2009.

[10] G. Obozinski, M. J. Wainwright, and M. I. Jordan, *Union support recovery in high-dimensional multivariate regression.* Technical report, Department of Statistics, UC Berkeley, August 2008.

[11] D. Paul and I. Johnstone, *Augmented sparse principal component analysis for high-dimensional data.* Technical report, UC Davis, January 2008.

[12] R. Tibshirani, *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, Series B, 58(1):267–288, 1996.

[13] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, *Algorithms for simultaneous sparse approximation.* Signal Processing, 86:572–602, April 2006.

[14] Kim Y., Kim J., and Y. Kim, *Blockwise sparse regression.* Statistica Sinica, 16(2), 2006.

[15] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society B, 1(68):49, 2006.

[16] P. Zhao, G. Rocha, and B. Yu, *Grouped and hierarchical model selection through composite absolute penalties.* Technical Report 703, Statistics Department, University of California, Berkeley, 2007.

Higher Criticism Thresholding for Optimal Feature Selection

David Donoho

(joint work with Jiashun Jin)

1. When useful features are rare and weak

Consider a two-class classification setting where we have a set of labeled training samples (Y_i, X_i), $i = 1, 2, \dots, n$. Each label $Y_i = 1$ if X_i comes from Class 1 and $Y_i = -1$ if X_i comes from Class 2, and each feature vector $X_i \in R^p$. For simplicity, we suppose that the training set contains equal numbers of samples from each of the two classes, and that the feature vector obeys $X_i \sim N(Y_i\mu, I_p)$, $i = 1, 2, \dots, n$, for an unknown mean contrast vector $\mu \in R^p$. Also, we suppose the feature covariance matrix is the identity matrix.

Formally our goal is to use the training data to design a classifier for use on fresh data. If we are given a new unlabelled feature vector X, we must then label it with a class prediction, i.e. attach a label $\hat{Y} = 1$ or $\hat{Y} = -1$. We hope that our predicted label $\hat{Y}$ is typically correct.

Following our papers, we consider the following *rare/weak feature model*, where the vector μ is nonzero in only a fraction ϵ of coordinates, and the nonzero coordinates of μ share a common amplitude μ_0. For $\tau = \sqrt{n}\mu_0$, let $RW(\epsilon, \tau; n, p)$ denote this model.

1.1. **Linking Rarity and Weakness to Number of Features.** We now adopt an *asymptotic* viewpoint, letting the number of features p be the driving problem size descriptor and for the purposes of calculation, we let p tend to infinity. Other problem parameters also depend on p as follows. Fixing parameters $(\beta, r) \in (0,1)^2$, let

$$\epsilon = \epsilon_p = p^{-\beta}, \qquad \tau = \tau_p = \sqrt{2r \log p}.$$

As $p \to \infty$, the useful features become increasingly rare; an asymptotically negligible fraction of the components in the vector Z. The parameters (β, r) describe the linkage between rareness and weakness of the entries in the parameter vector. The domain $(\beta, r) \in (0,1)^2$ will be seen to have an interesting two-phase structure; we call a depiction of this domain and its phases a *phase diagram.*

1.2. **Linking Number of Observations to Number of Features.** The phase diagram depends on the relationship between the number of features p and the number of study units n. Again, in our work it is convenient to make p the driving variable, and so $n = n_p$.

We can identify three regimes for the linkage between n and p: $n = n_p$ can have *no growth*, *slow growth*, or *regular growth.* Our labels for these regimes and their definitions go as follows:

Regime	Label	Definition
No Growth	(N)	$n_p = n_0$ for some constant n_0
Slow Growth	(S)	$n_p \to \infty$, but $n_p/p^\theta \to 0$, $\forall\, \theta > 0$
Regular Growth	(R)	$n_p = p^\theta$ for some $\theta \in (0,1)$

1.3. **Asymptotic Rare/Weak Model (ARW).** Combining the two linkages we have just discussed gives us the *asymptotic rare/weak model* $ARW(\beta, r, n_p)$. Our goal is to determine the underlying *phase structure.* For each linkage type n_p we seek to identify ranges of (β, r) where successful classification is possible and impossible, respectively.

2. Impossibility of Classification

Jin (2009) and, at this workshop, also Ingster, Pouet and Tsybakov (2009) show that in each of the three growth regimes, there is a curve $r = \rho^\star(\beta)$ $(\star = N, S, R)$ which partitions the β-r plane into two components: a *region of impossibility* below the curve and and *region of possibility* above it. In detail, define the *standard phase boundary* function

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

Define

$$\begin{aligned}\rho^N(\beta) &= \rho^N(\beta, n) = \frac{n}{n+1}\rho(\beta), \qquad 0 < \beta < 1,\\ \rho^S(\beta) &= \rho(\beta), \qquad 0 < \beta < 1,\\ \rho^R(\beta) &= (1-\theta)\rho(\beta/(1-\theta)), \qquad 0 < \beta < (1-\theta).\end{aligned}$$

Jin (2009) and Ingster, Pouet, Tsybakov (2009) suppose that we fix a growth regime n_p and fix a point (β, r) in the region 'below' the corresponding graph $(\beta, \rho^\star(\beta))$. Consider the sequence of problems $ARW(r, \beta, n_p)$ for increasing p and any sequence of classifier training methods, perhaps also dependent on p. The misclassification error rate of the resulting trained classifier → 1/2 as $p \to \infty$. In this region, the measurements are effectively non-informative, and random guessing does almost as well

3. Success by Higher Criticism Thresholding

Donoho and Jin (2008a) introduced a technique for feature selection in the high-dimensional $p > n$ sparse case. Donoho and Jin (2008b) showed that in the case of slow growth S, this method obtains is successful throughout the full interior of the complement of the impossibility region. Moreover, the possibility/impossibility dichotomy in the phase diagram can be further split into 3 interesting phases, in which the optimal feature selection threshold and the HC feature selection threshold have limiting false feature selection rates with some surprising properties. In particular, in Region I, we are surprised to see that optimal behavior requires very high false discovery rate (1) which accompanies a misclassification probability tending to zero!

Definition 3.1. Regions I,II, III. *The Possibility Region can be split into Regions I-III with the following interiors:*

I.: $\beta - 1/2 < r \leq \beta/3$ *and* $1/2 < \beta < 3/4$; $r > \rho^*(\beta)$.

II.: $\beta/3 < r \leq \beta$ *and* $1/2 < \beta < 1$; $r > \rho^*(\beta)$.

III.: $\beta < r < 1$ *and* $1/2 < \beta < 1$; $r > \rho^*(\beta)$.

See Figure 1.

Acknowledgements

The authors would like to thank the organizers for their efforts in creating a very successful workshop, and the staff at Oberwolfach for their usual very friendly and effective service.

References

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

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

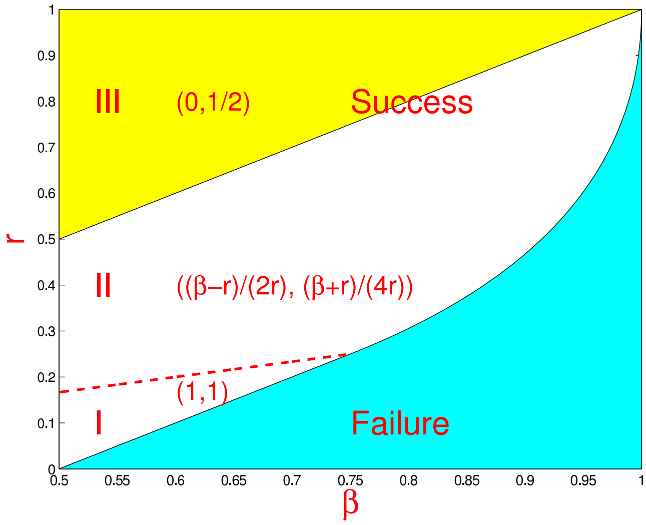

Figure 1. Phase diagram. The curve $r = \rho^*(\beta)$ splits the phase space into the impossibility region and the possibility region, and the latter further splits into three different regions I, II, III. Numbers in the brackets show limits of false feature discovery rate and local false feature discovery rate for both HCT feature selection and ideal threshold feature selection

[3] Donoho, D. & Jin, J. 2008b, *Feature Selection by Higher Criticism Thresholding: Optimal Phase Diagram.* Phil. Trans. Roy. Soc., in press. Eprint arxiv:math:0812.2263.

[4] Donoho, D. & Jin, J. 2009, *When useful features are rare and weak: HCT yields successful classification throughout the region of possibility.* Working manuscript.

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

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

[7] Ingster, Y., Pouet, C., & Tsybakov, A. B. 2009, *Classification of sparse high-dimensional vectors.* Manuscript.

[8] Ingster, Y., Pouet, C., & Tsybakov, A. 2009 b, *Sparse classification boundaries.* Eprint arxiv:math/09034807.

[9] Jin, J. 2009, *Impossibility of successful classification when useful features are rare and weak.* Proc. Nat. Acad. Sci., in press.

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

Exact and Robust Matrix Completion

Emmanuel J. Candès

(joint work with Benjamin Recht, Terence Tao and Yaniv Plan)

This talk considers a problem of great practical interest: the recovery of a data matrix from a sampling of its entries. In partially filled out surveys, for instance, we would like to infer the many missing entries. In the area of recommender systems, users submit ratings on a subset of entries in a database, and the vendor provides recommendations based on the user's preferences. Because users only rate a few items, we would like to infer their preference for unrated items (this is the famous Netflix problem). Formally, suppose that we observe $m \ll n^2$ entries selected uniformly at random from an $n \times n$ matrix. Can we complete the matrix and recover the entries that we have not seen?

We show that perhaps surprisingly, one can recover low-rank matrices exactly from what appear to be highly incomplete sets of sampled entries; that is, from a minimally sampled set of entries [4]. Further, perfect recovery is possible by solving a simple convex optimization program, namely, a convenient semidefinite program. A surprise is that our methods are optimal and succeed as soon as recovery is possible by any method whatsoever, no matter how intractable; this result hinges on powerful techniques in probability theory. Further, this talk introduces novel results showing that matrix completion is provably accurate even when the few observed entries are corrupted with a small amount of noise. A typical result is that one can recover an unknown $n \times n$ matrix of low rank r from just about $nr \log^2 n$ noisy samples with an error which is proportional to the noise level [1]. We present numerical results which complement our quantitative analysis and show that, in practice, nuclear norm minimization accurately fills in the many missing entries of large low-rank matrices from just a few noisy samples. Some analogies between matrix completion and compressed sensing are discussed throughout.

A typical result is as follows: suppose one observes m entries of large $n_1 \times n_2$ matrix M of rank r whose singular value decomposition is given by

$$M = \sum_{k \in [r]} \sigma_k u_k v_k^*, \tag{1}$$

in which $\sigma_1, \ldots, \sigma_r \geq 0$ are the singular values, and $u_1, \ldots, u_r \in \mathbb{R}^{n_1}$, $v_1, \ldots, v_r \in \mathbb{R}^{n_2}$ are the singular vectors. We propose recovering the unknowm matrix M by solving the convex optimization program

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega, \end{array} \tag{2}$$

where $\|X\|_*$ is the nuclear norm of X, i.e. the sum of the singular values of X, and Ω is the set of entries (i,j) which are observed. It is well known that (2) is a semidefinite program.

Now asssume that

$$\|u_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_1}, \quad \|v_k\|_{\ell_\infty} \leq \sqrt{\mu_B / n_2}, \tag{3}$$

for some $\mu_B \geq 1$, where the ℓ_∞ norm is of course defined by $\|x\|_{\ell_\infty} = \max_i |x_i|$. We think of μ_B as being small, e.g. $O(1)$, so that the singular vectors are not too spiky. In [4], it is proven that nuclear-norm minimization succeeds nearly as soon as recovery is possible by any method whatsoever.

Theorem 3. *[4] Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix of rank $r = O(1)$ obeying* (3) *and set $n := \max(n_1, n_2)$. Suppose we observe m entries of M with locations sampled uniformly at random. Then there is a positive numerical constant C such that if*

$$m \geq C\, \mu_B^4\, n \log^2 n, \tag{4}$$

then M is the unique solution to (2) *with probability at least $1 - n^{-3}$. In other words: with high probability, nuclear-norm minimization recovers all the entries of M with no error.*

When Ω is sampled at random, we show that to succeed, any method whatsoever needs at least on the order $O(n \log n)$ entries to succeed and, hence, (4) misses the information theoretic limit by at most a logarithmic factor.

We extend this result and show that $n_1 \times n_2$ matrices of arbirary rank r with singular vectors which are sufficiently spread can be recovered exactly via (2) provided that the number of samples m is on the order of $nr \log^6 n$ (where n is still the maximum between n_1 and n_2).

We also that *when perfect noiseless recovery occurs, then matrix completion is stable vis a vis perturbations* [1]. Suppose we observe

$$Y_{ij} = M_{ij} + Z_{ij}, \quad (i,j) \in \Omega, \tag{5}$$

where $\{Z_{ij} : (i,j) \in \Omega\}$ is a noise term which may be stochastic or deterministic (adversarial). All we assume is that $\sum_{(i,j)\in\Omega} Z_{ij}^2 \leq \delta^2$ for some $\delta > 0$. Then if $\hat{M}$ is the solution to the semidefinite program

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & \sum_{(i,j)\in\Omega} (X_{ij} - Y_{ij})^2 \leq \delta^2, \end{array} \tag{6}$$

we have that the error $\|\hat{M} - M\|_F$ is proportional to the noise level δ; when the noise level is small, the error is small.

Time permitting, we will rapidly discuss a very efficient algorithm based on iterative singular value thresholding, which can complete matrices with about a billion entries in a matter of minutes on a personal computer [2].

References

[1] E. J. Candès, and Y. Plan. *Matrix completion with noise.* Technical report, 2009. Preprint available at `http://arxiv.org/abs/0903.3131`.

[2] J-F. Cai, E. J. Candès, and Z. Shen. *A singular value thresholding algorithm for matrix completion.* Technical report, 2008. Preprint available at `http://arxiv.org/abs/0810.3286`.

[3] E. J. Candès and B. Recht. *Exact Matrix Completion via Convex Optimization.* To appear in *Found. of Comput. Math.*, 2008.

[4] E. J. Candès and T. Tao. *The power of convex relaxation: Near-optimal matrix completion.* Technical report, 2009. Submitted for publication and preprint available at `http://arxiv.org/abs/0903.1476`.

[5] R. Keshavan, S. Oh, and A. Montanari. *Matrix completion from a few entries.* Submitted to ISIT'09 and available at `arXiv:0901.3150`, 2009.

Spectral Methods for Surface Clustering

Ery Arias-Castro

(joint work with Gilad Lerman and Guangliang Chen)

Traditional methods for clustering, such as k-Means or Gaussian Mixture Models, assume that each cluster is generated by sampling points in the vicinity of a centroid, that is a point in space. The resulting clusters are ellipsoidal, and in particular full-dimensional. In a number of modern applications, however, the data seems to cluster near low-dimensional structures or surfaces. When the underlying low-dimensional surfaces are assumed to be affine subspaces, the problem of clustering and in general learning those structures is termed *Linear Hybrid Modeling*, which was recently featured in the *SIAM Review* [14], where the editor referred to it as "a very 'hot' area of research". Indeed, it applies to many real-life situations, such as motion segmentation in computer vision, hybrid linear representation of images, classification of face images, and temporal segmentation of video sequences [25, 14, 6]. Other settings call for modeling each cluster with a multilinear surface, for example, motion segmentation with multiple views using initial feature vectors [12, 18, 24, 22] and wearable action recognition [27]. In this paper we consider the more general situation where the underlying surfaces are nonparametric, which we refer to as *surface learning* as in [10], or to be more specific, *surface clustering*. This situation arises in a number of modern applications, e.g., motion segmentation without feature vectors [8] and the analysis of the galaxy distribution in Astrophysics, where structures such as filaments, sheets and spherical clusters are of interest [23, 15].

A number of approaches to surface clustering have been suggested, involving estimation of local characteristics such as dimensionality and density [7, 13], sometimes combined together in a global energy to be optimized, often with an EM-type algorithm [11, 10]; sometimes combining different off-the-shelf algorithms, such as ISOMAP and MDS [21]. Spectral methods seem to be the most popular. Originally based on pairwise affinities [16], more recent methods are now based on multi-way affinities, to better capture the actual complexity of the data [2, 9, 19, 1, 6].

Though the literature is growing rapidly, few papers rigorously analyze the performance of these algorithms, even in a simple mathematical model. In their paper [16], the authors introduce their method and outline a strategy to analyze it. However, the probabilistic analysis is not addressed. In [26], spectral clustering is taken to its empirical process limit as the number of points increases. Though this provides insight on what spectral clustering is estimating, there is no result on performance. The same comment applies to [17]. In [4], the number of clusters is estimated, and conditions are provided under which the estimation is consistent. Again, there is no result on performance. Our analysis is closer in spirit to the work [5] in the context of hybrid linear modeling, which is inspired by the strategy outlined in [16].

Our contribution is two-fold. First, we provide theoretical guaranties for the pairwise spectral clustering technique in the form described in [16]. We carry out their program under a fairly general generative model for surface learning. Second, we introduce a new multi-way spectral clustering method based on local linear (or higher order) approximations, for which we provide theoretical guaranties as well. In contrast with the pairwise clustering, this method is able to take advantage of the smoothness of the underlying surfaces. We discuss the choice of different parameters, including the scale(s) and the number of clusters. We also address the issue of outliers and show that with simple modifications, these spectral methods are highly robust, in fact able to accurately cluster within logarithmic factors of what is needed in [3] to merely detect the presence of those clusters. We also discuss the computational aspect of such spectral clustering techniq! ues. Assuming that the ambient dimension is not too large, a direct implementation of the pairwise algorithm is quadratic in the number of points, since it is based on all pairwise distances. Effectively, only a small fraction of those pairwise distances are significantly different from zero, so that for a given point only the pairwise distances to the closest neighbors need to be computed. We show that implementing k-nearest-neighbors techniques enables to substantially speed up the computations while maintaining a comparable clustering performance. The same approach works in the case of a multi-way affinity. We perform numerical experiments illustrating the theory.

References

[1] S. Agarwal, K. Branson, and S. Belongie, *Higher order learning with graphs.* In Proceedings of the 23rd International Conference on Machine learning, volume 148, pages 17–24, 2006.

[2] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie, *Beyond pairwise clustering.* In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 838–845, 2005.

[3] E. Arias-Castro, D. L. Donoho, X. Huo, and C. A. Tovey, *Connect the dots: how many random points can a regular curve pass through?* Adv. in Appl. Probab., 37(3):571–603, 2005.

[4] G. Biau, B. Cadre, and B. Pelletier, *A graph-based estimator of the number of clusters.* ESAIM Probab. Stat., 11:272–280, 2007.

[5] G. Chen and G. Lerman, *Foundations of a multi-way spectral clustering framework for hybrid linear modeling.* Submitted to the Journal of Foundations of Computational Mathematics. Available from `http://arxiv.org/abs/0810.3724v1`.

[6] G. Chen and G. Lerman, *Spectral curvature clustering (SCC).* To appear in Int. J. Comput. Vision.

[7] A. Gionis, A. Hinneburg, S. Papadimitriou, and P. Tsaparas, *Dimension induced clustering.* In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 51–60, New York, NY, USA, 2005. ACM.

[8] A. Goh and R. Vidal, *Clustering and dimensionality reduction on Riemannian manifolds.* IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pages 1–7.

[9] V. Govindu, *A tensor decomposition for geometric grouping and segmentation.* In *CVPR*, volume 1, pages 1150–1157, June 2005.

[10] Q. Guo, H. Li, W. Chen, I.-F. Shen, and J. Parkkinen, *Manifold clustering via energy minimization.* In *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications*, pages 375–380, Washington, DC, USA, 2007. IEEE Computer Society.

[11] G. Haro, G. Randall, and G. Sapiro, *Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds.* Neural Information Processing Systems, 2006.

[12] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521623049, 2000.

[13] D. Kushnir, M. Galun, and A. Brandt, *Fast multiscale clustering and manifold identification.* Pattern Recogn., 39(10):1876–1891, 2006.

[14] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. *Estimation of subspace arrangements with applications in modeling and segmenting mixed data.* SIAM Review, 50(3):413–458, 2008.

[15] V. Martínez and E. Saar. *Statistics of the Galaxy Distribution.* Chapman and Hall/CRC press, Boca Raton, 2002.

[16] A. Ng, M. Jordan, and Y. Weiss. *On spectral clustering: Analysis and an algorithm.* In Advances in Neural Information Processing Systems, volume 14, pages 849–856, 2002.

[17] B. Pelletier and P. Pudlo. *Strong consistency of spectral clustering on level sets.* Available from `http://www.math.univ-montp2.fr/~pelletier/publications.html`.

[18] S. Rao, A. Yang, S. Sastry, and Y. Ma. *Robust algebraic segmentation of mixed rigid-body and planar motions.* Submitted to International Journal of Computer Vision, 2008.

[19] A. Shashua, R. Zass, and T. Hazan. *Multi-way clustering using supersymmetric non-negative tensor factorization.* In ECCV06, volume IV, pages 595–608, 2006.

[20] C. Stoughton, R. Lupton, M. Bernardi, M. Blanton, S. Burles, F. Castander, A. Connolly, D. Eisenstein, J. Frieman, G. Hennessy, et al. *Sloan Digital Sky Survey: Early Data Release.* The Astronomical Journal, 123(1):485548, 2002.

[21] R. Souvenir and R. Pless. *Manifold clustering.* volume 1, pages 648–653 Vol. 1, 2005.

[22] R. Tron and R. Vidal. *A benchmark for the comparison of 3-d motion segmentation algorithms.* In CVPR, 2007.

[23] R. Valdarnini. *Detection of non-random patterns in cosmological gravitational clustering.* Astronomy & Astrophysics, 366:376–386, 2001.

[24] R. Vidal and R. Hartley. *Three-view multibody structure from motion.* IEEE Trans. Pattern Anal. Mach. Intell., 30(2):214–227, 2008.

[25] R. Vidal, Y. Ma, and S. Sastry. *Generalized principal component analysis (GPCA).* IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(12), 2005.

[26] U. von Luxburg, M. Belkin, and O. Bousquet. *Consistency of spectral clustering.* The Annals of Statistics, 36(2):555–586, 2008.

[27] A. Yang, R. Jafari, P. Kuryloski, S. Iyengar, S. S. Sastry, and R. Bajcsy. *Distributed segmentation and classification of human actions using a wearable motion sensor network.* Technical Report UCB/EECS-2007-143, EECS Department, University of California, Berkeley, Dec 2007.

ROC curve estimation and optimization

Nicolas Vayatis
(joint work with Stéphan Clémençon)

1. Introduction

Since their introduction in the sixties, Receiver Operating Characteristic (ROC) curves have been extensively used in signal detection, medical diagnosis and credit risk screening as a visual tool for performance assessment. Following the statistical learning approach, we propose to use ROC curves as a primary target for optimization procedures in the context of ranking and scoring applications. We develop various approaches for the approximation and estimation of the optimal ROC curve. We explore the statistical properties of estimators based on summaries of ROC curves and sketch algorithmic schemes for practical implementation.

2. Definitions and notations

We consider classification data obtained from sampling a random pair (X, Y) where $X \in \mathbb{R}^d$ is a vector of covariates and $Y \in \{-1, +1\}$ is a binary label. We denote the regression function by $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$. We aim at ranking the data rather than just classifying them into positive and negative labels. That is, we aim at recovering the order induced by the regression function η rather than the single level set $\{x : \eta(x) > 1/2\}$. This problem is known as the *bipartite ranking problem.* The candidate predictors in this problem are real-valued functions $s : \mathbb{R}^d \to \mathbb{R}$ called scoring rules and their performance can be measured by monitoring simultaneously two quantities called the *false positive rate* and the *true positive rate.* For a scoring rule s and a threshold $t \in \mathbb{R}$, we define the false positive rate by $\alpha(s, t) = \mathbb{P}\{s(X) \geq t \mid Y = -1\}$ and the true positive rate by

$\beta(s,t) = \mathbb{P}\{s(X) \geq t \mid Y = +1\}$. The ROC curve of a particular scoring rule is the parametric curve obtained in the plane (α, β) when t goes from $-\infty$ to $+\infty$. Interestingly, the ROC curve is invariant by strictly increasing transforms of the scoring rule s and the optimal ROC curve (the one which, for each α, dominates all the others) corresponds to the regression function η by a simple application of Neyman-Pearson's lemma.

3. Summaries of ROC curves

In order to estimate optimal scoring rules, the inference principles proposed in the literature rely on functionals built from the ROC curve. We call these functionals *summaries.*

The AUC criterion. The most popular example is the Area Under an ROC curve (AUC). For any scoring rule $s : \mathbb{R}^d \to \mathbb{R}$ with non-constant parts, the AUC can easily be interpreted as the following probability:

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X) > s(X') \mid (Y, Y') = (+1, -1)\}$$

where (X, Y), (X', Y') are i.i.d.. This expression shows that maximizing the AUC is equivalent to minimizing a classification error in a classification problem with input (X, X') and output $Y - Y'$. Empirical risk minimization in this case leads to the analysis of U-processes. Consistency, rates-of-convergence type results and fast rates of convergence can be derived in this setup ([1]).

The local AUC. However, the AUC criterion does not reflect performance in restricted parts of the input space since it weights uniformly discordant pairs of observations. In order to focus on best instances, we propose to localize the AUC criterion. We need to introduce as a parameter to control the localization the rate $u \in (0,1)$ of "best" X's. We can then redefine the parametrization of the ROC curve using the new parameter u. The level of truncation of the AUC corresponding to a rate u of best instances is specified by the control line $u = p\beta + (1-p)\alpha$ in the ROC space (α, β). We show ([2]) that a consistent criterion for finding the optimal ranking over the set of best instances is provided by the sum of the truncated AUC at the level defined by u plus the term $(1-\alpha)\beta$.

General summaries. It is interesting to notice that both the AUC, and the local AUC, as well as other functionals introduced in the machine learning and information retrieval literature, such as the p-norm push (Rudin - JMLR, 2006), or the Discounted Cumulative Gain (Cossock and Zhang - COLT 2006), can be expressed as *conditional linear rank statistics* of the form:

$$W_n(s) = \sum_{i=1}^{n} \mathbb{I}\{Y_i = 1\} \; \Phi\left(\frac{\mathrm{rank}(s(X_i))}{n+1}\right)$$

where $\Phi : [0,1] \to [0,1]$ is called the score-generating function. For maximizers of such functionals, consistency results can be proved by using Hájek's projection and providing a uniform control of the remainder term in the decomposition ([4]). However, a general theory for R-processes needs to be further developed.

4. Algorithmic insights

Optimizing the ROC curve requires to simultaneously approximate the curve and estimate its points. This can be carried out along two directions: (i) building partitions of the input space, (ii) taking partitions of $[0, 1]$ in the α-axis of the ROC space. The first approach can be illustrated through dedicated versions of decision trees or histogram rules that we have studied ([5, 6]). The second approach can be implemented by building on the following observation: optimal scoring rules can be represented from the collection of level sets of the regression function. For instance, for the regression function, we have the identity:

$$\eta(x) = \mathrm{E}\left(\mathbb{I}\{\eta(x) > U\}\right)$$

where U is a uniform random variable on $[0, 1]$. Hence, finding an optimal scoring rule in the sense of the ROC curve may be viewed as dealing with a 'continuum' of classification problems with asymmetric costs where the targets are the level sets. For practical considerations, discretization of the previous formula is a key point (fixed vs. adaptive partitions of $[0, 1]$). Once the level is fixed, we propose to use an empirical minimum-volume set estimation in order to learn each of these level sets and we provide rates of convergence with which a point of the optimal ROC curve can be recovered according to this principle. From the resulting classifiers and their related empirical errors, we show ([3]) how to build a linear-by-part estimate of the optimal ROC curve and a quasi-optimal piecewise constant scoring rule. Rate bounds in terms of sup-norm in the ROC space for these procedures are also established.

References

[1] S. Clémençon, G. Lugosi, and N. Vayatis (2008), *Ranking and empirical risk minimization of U-statistics.* The Annals of Statistics, 36: 844–874.

[2] S. Clémençon and N. Vayatis (2007), *Ranking the best instances.* Journal of Machine Learning Research, 8(Dec):2671-2699.

[3] S. Clémençon and N. Vayatis (2008a), *Overlaying classifiers: a practical approach for optimal ranking.* NIPS '08: Proceedings of the 2008 Conference on Advances in Neural Information Processing Systems. Vancouver, Canada.

[4] S. Clémençon and N. Vayatis (2008b), *Empirical performance maximization for linear rank statistics.* NIPS '08: Proceedings of the 2008 Conference on Advances in Neural Information Processing Systems. Vancouver, Canada.

[5] S. Clémençon and N. Vayatis (2008b), *Tree-based ranking methods.* Technical Report hal-00268068, HAL.

[6] S. Clémençon and N. Vayatis (2009), *On partitioning rules for bipartite ranking.* Journal of Machine Learning Research - Proceedings of AISTATS '09.

Sparse Recovery by Aggregation and Langevin Monte-Carlo

Arnak Dalalyan
(joint work with A.B. Tsybakov)

The aim of the presented work is to put forward the usefulness of the exponentially weighted aggregate in the problem of estimation of a high dimensional parameter-vector under sparsity scenario. More precisely, we first consider the model of regression

$$Y_i = f(Z_i) + \xi_i, \qquad i = 1, \ldots, n,$$

where $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ are observed, f is the unknown regression function and $\xi_1, \ldots, \xi_n$ are random variables assumed to be independent identically distributed with zero mean and finite variance σ^2. The regressors Z_i are assumed to be deterministic. Furthermore, we assume that we have at our disposal a parametric family $\mathcal{F}_\Lambda$ of functions f_λ, $\lambda \in \Lambda \subset \mathbb{R}^M$, for some $M \in \mathbb{N}$, that may be used for estimating the regression function f.

For every prior distribution π over Λ and for every $\beta > 0$, the exponentially weighted aggregate (EWA) is defined by

$$\hat{f}_n = \frac{\int_\Lambda f_\lambda \exp(-\beta^{-1}\mathtt{RSS}(\lambda))\,\pi(d\lambda)}{\int_\Lambda \exp(-\beta^{-1}\mathtt{RSS}(\lambda))\,\pi(d\lambda)},$$

where the residual sum of squares is $\mathtt{RSS}(\lambda) = \sum_{i=1}^n (Y_i - f_\lambda(Z_i))^2$. We prove an oracle inequality for this estimator stating that under mild assumptions, for every $\beta \geq 4\sigma^2$, it holds

$$\mathbf{E}[\|\hat{f} - f\|_n^2] \leq \inf_p \Big(\int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) + \frac{\beta \mathcal{K}(p,\pi)}{n} \Big),$$

where $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(Z_i)^2$ for every function g, the infimum is taken over all probability distributions p satisfying $\int_\Lambda \|f_\lambda\|_n^2 p(d\lambda) < \infty$ and $\mathcal{K}(p,\pi)$ stands for the Kullback-Leibler divergence between p and the prior distribution π.

We then apply the obtained inequality to the linear model

$$Y_i = x_i^\top \lambda^* + \delta_i + \xi_i, \qquad i = 1, \ldots, n,$$

where (x_i, Y_i) are observed, $\lambda^* \in \mathbb{R}^M$ is the unknown parameter-vector, δ_i is a sequence of deterministic errors while ξ_i is a sequence of random errors satisfying the same assumptions as above. We place ourselves under the sparsity scenario, that is we allow M to be of the same order as or even larger than the sample size n, but we assume that only a small number of coordinates of λ^* are different from zero. We choose a prior distribution π that is, roughly speaking, the product of scaled univariate Student distributions with 3 degrees of freedom. The EWA $\hat{\lambda}_n$ based on this prior and a properly chosen scaling parameter is shown to satisfy the inequality

$$\mathbf{E}[\|\hat{\lambda}_n - \lambda^*\|_X^2] \leq \frac{\|\delta\|_2^2 + 4\beta \sum_{j=1}^M \log(1 + |\lambda_j^*/\sigma|\sqrt{nM})}{n},$$

where, by definition, $\|v\|_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i^\top v)^2$. A remarkable point in this result is that it is obtained under no assumption on the covariates $\{x_i\}$ and that the sum in the right hand side of the inequality is proportional, up to a logarithmic factor, to the number of non-zero coordinates of λ^*, which is expected to be small in view of the sparsity assumption. The rigorous statements of these results can be found in [1, 2].

Finally, we introduce a version of Langevin Monte-Carlo procedure, that allows us to efficiently compute an approximation to the EWA even for large values of M. We illustrate the quality of estimation of large vectors by the EWA and the efficiency of the Langevin Monte-Carlo approximation through numerical simulations.

References

[1] Dalalyan, A.S. and Tsybakov, A.B.: *Aggregation by exponential weighting, sharp oracle inequalities and sparsity.* Machine Learning, **72** (2008), no. 1-2, 39– 61.

[2] Dalalyan, A.S. and Tsybakov, A.B.: *Sparse Regression Learning by Aggregation and Langevin Monte-Carlo.* Submitted.

On an incomplete aggregation of hard thresholding estimates

Yuri Golubev

This paper focuses on recovering an unknown vector $\theta \in \mathbb{R}^n$ from the noisy data

$$Y_i = \theta_i + \sigma\xi_i, \quad i = 1, \ldots, n,$$

where ξ_i are Gaussian i.i.d. random variables $\mathcal{N}(0,1)$. In order to estimate θ, we use the hard thresholding method

$$\hat{\theta}_i(t,Y) = Y_i \mathbf{1}\big\{|Y_i| \geq t\big\}$$

and our goal is to choose the threshold t based on the data at hand. The motivation of the hard thresholding technique is based on the hypotheses that the underlying vector θ is sparse. Such vectors typically arise in statistical problems related to the wavelets technique, see for instance Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995).

In this paper, the thresholds are chosen with the help of the principle of the empirical risk minimization, i.e.,

$$\hat{t} = \arg\min_{t\in\mathcal{T}} \Big\{ \|Y - \hat{\theta}(t,Y)\|^2 + Pen\big[\|\hat{\theta}(t,Y)\|_0, t\big] \Big\},$$

where $\mathcal{T}$ is a subset in $\mathbb{R}^+$, $\|\cdot\|$ is the usual Eucledian norm, $\|x\|_0 = \sum_{i=1}^n \mathbf{1}\{|x_i| > 0\}$, and $Pen\big[\cdot,\cdot\big]$ is a penalty function. Generally speaking, the main goal in the empirical risk minimization is to find a universal penalty that minimizes the risk of $\hat{\theta}\big(\hat{t}, Y\big)$

$$R(\theta, Pen) = \mathbf{E}\sum_{i=1}^{n}\big[\theta_i - \hat{\theta}_i\big(\hat{t}, Y\big)\big]^2$$

uniformly in $\theta \in \mathbb{R}^n$. To see how this method works, let us start with the classical situation, where $\mathcal{T} = \mathbb{R}^+$. In statistical literature, this case is often called *model selection* (see e.g., Birgé, L. and Massart, P. (2007) for its concise history).

In order to bound from above the risk of $\hat{\theta}(\hat{t}, Y)$, let us introduce some additional notations. For a vector $x \in \mathbb{R}^n$ we denote by $x_{(\cdot)}$ the decreasing permutation of its components, i.e. $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)}$. Let

$$pen_\alpha(k) \overset{\text{def}}{=} (2k+1)\log\frac{2Q_\alpha n}{2k+1}, \quad \text{where } Q_\alpha = \left(1 + \frac{1}{\alpha}\right)\exp(2+\epsilon) \quad \text{with } \epsilon > 0.$$

We begin with a rough result in the spirit of Birgé, L. and Massart, P. (2007).

Theorem 4. *Let $\mathcal{T} = \mathbb{R}^+$ and $Pen_\alpha(x) = (1+\alpha)\sigma^2 pen_\alpha(x)$, then for any $\epsilon > 0$, uniformly in $\theta \in \mathbb{R}^n$*

$$R(\theta, Pen_\alpha) \leq \left(1 + \frac{1}{\alpha}\right)\mathcal{R}(\theta, Pen_\alpha) + \frac{C\sigma^2(1+\alpha)^2}{\alpha\sqrt{\epsilon}},$$

where $\mathcal{R}(\theta, Pen_\alpha) = \min_k\Big\{\sum_{i=k+1}^n \theta_{(i)}^2 + Pen_\alpha(k)\Big\}$ and here and in the sequel $C > 0$ stands for a generic positive constant.

This theorem predicts the following properties of the empirical risk minimization: *it blows up rapidly as $\alpha \to 0$, the nearly optimal penalty corresponds to $\alpha \approx 1$, the risk of this method (with the optimal penalty) over the set $\mathbb{S}_{\gamma_n} = \{\theta : \|\theta\|_0 \leq \gamma_n n\}$ is asymptotically fourfold the minimax risk (see Abramovich, F., Benjamini, Y., Donoho, D. and Johstone, I., (2006)).* Unfortunately, these conclusions are far from statistical practice since they result from the imprecise upper bound.

The next result shows that the upper bound from Theorem 1 may be improved. In what follows it is assumed that

- $\mathcal{T} = \{t_1, t_2, \ldots, t_M\}$, where $M = n/\big[\exp(1)\sqrt{\log(n)}\big]$ and t_k are defined by

$$t_k = \sigma\sqrt{2\log\frac{n}{1 + (k-1)\sqrt{\log(n)}}}$$

- $Pen_\alpha(k,t) = (1+\alpha)\sigma^2 pen_{\alpha/2}\big(k\big) + \alpha n\big(t^2/2 + \sigma^2\big)\exp[-t^2/(2\sigma^2)]$ with some $\alpha > 0$.

Theorem 5. *Uniformly in $\theta \in \mathbb{Q}^n$ and $t \in \mathcal{T}$*

$$\begin{aligned} R(\theta, Pen_\alpha) \leq& \mathcal{R}_0(\theta, t) + \sigma^2\#\left\{i : |\theta_i| \geq \frac{\sigma}{\sqrt{\log(n)}}\right\} \\ &+ \frac{C\mathcal{R}_0(\theta, t)}{\alpha^{3/2}\epsilon^{1/4}}\left[1 + \log_+\frac{C\sqrt{\epsilon}\alpha^3 n\sigma^2}{\mathcal{R}_0(\theta, t)}\right]^{-1/4}, \end{aligned} \tag{1}$$

where $\mathcal{R}_0(\theta,t)=\sum_{i=1}^n\theta_i^2\mathbf{P}\big\{|Y_i|\le t\big\}+Pen_\alpha\Big[\sum_{i=1}^n\mathbf{P}\big\{|Y_i|>t\big\},t\Big]$.

Statistical sense of this theorem is rather transparent. If the underlying vector is sparse, then $\mathcal{R}_0(\theta,t)\ll n\sigma^2$ for reasonable thresholds t. This means that $\mathcal{R}_0(\theta,t)$ is the main term at the right-side in (1) and its minimum in α is attained at $\alpha\approx 0$. Notice also that this upper bound blows up, but in contrast to Theorem 4, this effect takes place in the second order term. Since there is no blowup in the main term, one can prove that $\hat\theta(\hat t,Y)$ is the nearly asymptotically minimax estimator over $\mathbb{S}_{\gamma_n}$.

References

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

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

[3] Donoho D., Johnstone I., Kerkyacharian G., and Picard D., (1995), *Wavelet shrinkage: Asymtopia?.* Journal of the Royal Statistical Society, Series B, **57**, 301–369.

Asymptotic efficiency of simple decisions for the compound decision problem and estimating the mean of high value observations

Yaacov Ritov

(joint work with E. Greenstein and J. Park)

Let $\mathcal{F}=\{F_\mu:\mu\in\mathcal{M}\}$ be a parameterized family of distributions. Let $Y_1,Y_2\dots$ be a sequence of independent random variables, where Y_i takes value in some space $\mathcal{Y}$, and $Y_i\sim F_{\mu_i}$, $i=1,2,\dots$. For each n, we suppose that the sequence $\mu_{1:n}$ is known up to a permutation, where for any sequence $x=(x_1,x_2,\dots)$ we denote the sub-sequence $x_s,\dots,x_t$ by $x_{s:t}$. We denote by $\boldsymbol{\mu}=\boldsymbol{\mu}_n$ the set $\{\mu_1,\dots,\mu_n\}$, i.e., $\boldsymbol{\mu}$ is $\mu_{1:n}$ without any order information. We consider in this note the problem of estimating $\mu_{1:n}$ by $\hat\mu_{1:n}$ under the loss $\sum_{i=1}^n(\hat\mu_i-\mu_i)^2$, where $\hat\mu_{1:n}=\Delta(Y_{1:n})$. We assume that the family $\mathcal{F}$ is dominated by a measure ν, and denote the corresponding densities simply by $f_i=f_{\mu_i}$, $i=1,\dots,n$. The important example is, as usual, $F_{\mu_i}=N(\mu_i,1)$.

Let $\mathcal{D}^S=\mathcal{D}_n^S$ be the set of all *simple symmetric decision functions* Δ, that is, all Δ such that $\Delta(Y_{1:n})=(\delta(Y_1),\dots,\delta(Y_n))$, for some function $\delta:\mathcal{Y}\to\mathcal{M}$. In particular, the best simple symmetric function is denoted by $\Delta_{\boldsymbol{\mu}}^S=(\delta_{\boldsymbol{\mu}}^S(Y_1),\dots,\delta_{\boldsymbol{\mu}}^S(Y_n))$:

$$\Delta_{\boldsymbol{\mu}}^S=\operatorname*{arg\,min}_{\Delta\in\mathcal{D}_n^S}\mathrm{E}\,||\Delta-\mu_{1:n}||^2,$$

and denote

$$r_n^S=\mathrm{E}\,||\Delta_{\boldsymbol{\mu}}^S(Y_{1:n})-\mu_{1:n}||^2,$$

where, as usual, $\|a_{1:n}\|^2=\sum_{i=1}^n a_i^2$.

The class of simple rules may be considered too restrictive. Since the μs are known up to a permutation, the problem seems to be of matching the Ys to the μs. Thus, if $Y_i \sim N(\mu_i, 1)$, and $n = 2$, a reasonable decision would make $\hat{\mu}_1$ closer to $\mu_1 \wedge \mu_2$ as Y_2 gets larger. The simple rule clearly remains inefficient if the μs are well separated, and generally speaking, a bigger class of decision rules may be needed to obtain efficiency. However, given the natural invariance of the problem, it makes sense to be restricted to the class $\mathcal{D}^{PI} = \mathcal{D}_n^{PI}$ of all permutation invariant decision functions, i.e, functions Δ that satisfy for any permutation π and any $(Y_1, \dots, Y_n)$:

$$\Delta(Y_1, \dots, Y_n) = (\hat{\mu}_1, \dots, \hat{\mu}_n) \quad \Longleftrightarrow \quad \Delta(Y_{\pi(1)}, \dots, Y_{\pi(n)}) = (\hat{\mu}_{\pi(1)}, \dots, \hat{\mu}_{\pi(n)}).$$

Let

$$\Delta_{\boldsymbol{\mu}}^{PI} = \operatorname*{arg\,min}_{\Delta \in \mathcal{D}^{PI}} \mathrm{E}\, ||\Delta(Y^n) - \mu_{1:n}||^2$$

be the optimal permutation invariant rule under $\boldsymbol{\mu}$, and denote its risk by

$$r_n^{PI} = E||\Delta_{\boldsymbol{\mu}}^{PI}(Y_{1:n}) - \mu_{1:n}||^2.$$

Obviously $\mathcal{D}^S \subset \mathcal{D}^{PI}$, and whence $r_n^S \geq r_n^{PI}$. Still, 'folklore', theorems in the spirit of De Finetti, and results like Hannan and Robbins (1955), imply that asymptotically (as $n \to \infty$) $\Delta_{\mu^n}^{PI}$ and $\Delta_{\mu^n}^{S}$ will have 'similar' mean risks: $r_n^S - r_n^{PI} = o(n)$. Our main result establishes conditions that imply the stronger claim, $r_n^S - r_n^{PI} = O(1)$.

An asymptotic equivalence as above implies, that when we confine ourselves to the class of permutation invariant procedures, we may further restrict ourselves to the class of simple symmetric procedures, as is usually done in the standard analysis of compound decision problems. The later class is smaller and simpler.

The motivation for this paper stems from the way the notion of oracle is used in some sparse estimation problems. Consider two oracles, both *know* the value of $\boldsymbol{\mu}$. Oracle I is restricted to use only a procedure from the class $\mathcal{D}^{PI}$, while Oracle II is further restricted to use procedures from $\mathcal{D}^S$. Obviously Oracle I has an advantage, our results quantify this advantage and show that it is asymptotically negligible. Furthermore, starting with Robbins (1951) various oracle-inequalities were obtained showing that one can achieve nearly the risk of Oracle II, by a 'legitimate' statistical procedure. See, e.g., the survey Zhang (2003), for oracle-inequalities regarding the difference in risks. See also Brown and Greenshtein (2007), and Jiang and Zhang (2007) for oracle inequalities regarding the ratio of the risks. However, Oracle II is limited, and hence, these claims may seem to be too weak. Our equivalence results, extend many of those oracle inequalities to be valid also with respect to Oracle I. We needed a stronger result than the usual objective that the mean risks are equal up to $o(1)$ difference. Many of the above mentioned recent applications of the compound decision notion are about sparse situations when most of the μs are in fact 0, the mean risk is $o(1)$, and the only interest is in total risk.

As an example of the type of results assume that μ can get one of two values which we denote by $\{0,1\}$. To simplify notation we denote the two densities by f_0 and f_1.

Theorem 6. *Suppose that either of the following two conditions holds:*

(i) $f_{1-\mu}(Y_1)/f_\mu(Y_1)$ *has a finite variance under* $\mu \in \{0,1\}$.

(ii) $\sum_{i=1}^n \mu_i/n \to \gamma \in (0,1)$, *and* $f_{1-\mu}(Y_1)/f_\mu(Y_1)$ *has a finite variance under* $\mu = 0$.

Then $\mathrm{E}_{\boldsymbol{\mu}} \|\hat{\mu}^S - \hat{\mu}^{PI}\|^2 = O(1)$.

Suppose now that $Y_i \sim \mathcal{N}(\mu_i, 1)$. We assume that the vector $\mu = (\mu_1, \dots, \mu_n)$ is sparse, in the sense that most of the μ_i's are 0. A few of the Y_is are selected for a further investigation, and they should correspond to those with relatively large value of μ_i. Without any auxiliary information, the natural (and in fact the only conceivable) selection procedure is selecting those items with large value of Y_i. We want to investigate the total amount of signal in the selected lot. Formally, we study the estimation of

$$S_C = \sum \mu_i \mathbf{I}(Y_i > C),$$

for some given fixed C. We consider this problem as compound decision problem wand suggest an estimator.

Theorem 7. *Suppose*

$$f(C) < \kappa_1 \varphi(C),$$
$$|f''(C)| < \kappa_2 |\varphi''(C)|,$$

where the bounds κ_i $i = 1, 2$, *are uniform in* n. *Then*

$$\hat{S_C} = \eta + O_p(n^{\frac{3}{5}}[\varphi(C)(|\varphi''(C)|)^{\frac{1}{2}}]^{\frac{2}{5}}) = \eta + O_p(C[n\varphi(C)]^{\frac{3}{5}})$$
$$\hat{S_C} = S_C + O_p(C[n\varphi(C)]^{\frac{3}{5}})$$

On aggregation/selection of estimators

Alexander Goldenshluger

The subject of this paper is the problem of aggregating estimators from a given collection.

Consider the Gaussian white noise model

$$(1) \qquad Y_\varepsilon(\mathrm{d}t) = f(t)\mathrm{d}t + \varepsilon W(\mathrm{d}t), \quad t = (t_1, \dots, t_d) \in \mathcal{D}_0 = [0,1]^d,$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown function, $\varepsilon \in (0,1)$, and W is the standard Wiener process in $\mathbb{R}^d$. Let $\Theta \subset \mathbb{R}^N$ be a compact set, and assume that we are given a parameterized family of estimators $\mathcal{F}_\Theta = \{f_\theta,\ \theta \in \Theta\}$ of f. The objective is, using the observation $\mathcal{Y}_\varepsilon = \{Y_\varepsilon(t), t \in \mathcal{D}_0\}$, to select a single estimator from $\mathcal{F}_\Theta$ with the risk that is as close as possible to the risk of the best estimator in the family $\mathcal{F}_\Theta$. We refer to the outlined setup as the *aggregation problem.* Aggregation is a common approach to construction of nonparametric adaptive estimators; this fact motivates consideration of aggregation problems.

Let $\tilde{f}$ be an estimator of f based on the observation $\mathcal{Y}_\varepsilon$. We measure accuracy of $\tilde{f}$ by its $\mathbb{L}_p$–risk

$$\mathcal{R}_p[\tilde{f}; f] := \mathbb{E}_f \|\tilde{f} - f\|_p, \quad 1 \le p \le \infty,$$

where $\mathbb{E}_f$ is the expectation with respect to the probability measure $\mathbb{P}_f$ of observation $\mathcal{Y}_\varepsilon$ under model (1), and $\|\cdot\|_p$ is the standard $\mathbb{L}_p$–norm on $\mathcal{D}_0$. We want to propose a measurable choice, say $\hat{f} = f_{\hat{\theta}}$, from collection $\mathcal{F}_\Theta$ such that the following $\mathbb{L}_p$*–risk oracle inequality* holds:

$$\mathcal{R}_p[\hat{f}; f] \le C \inf_{\theta \in \Theta} \mathcal{R}_p[f_\theta; f] + r_\varepsilon \tag{2}$$

for all f from a ”large” functional class.

Here C is a constant independent of f and ε, and r_ε is a remainder term that does not depend on f.

We propose a general aggregation scheme that is universal in the following sense: (i) it applies to families of arbitrary estimators; (ii) it can be easily extended to different models; (iii) it can be used for a wide variety of global risk measures. Although the main results of this paper pertain to the MS aggregation setup, Gaussian white noise model and $\mathbb{L}_p$–risks, similar results can be easily established for other models and global risk measures. We illustrate universality of the suggested procedure by applying it to convex aggregation and to the problem of estimating a normal mean vector.

Our aggregation method is based on comparison of empirical estimates of certain regular linear functionals with estimates induced by the family $\mathcal{F}_\Theta$.

A closely related idea that a nonparametric function estimator is ”good” if its integrals over cubes ”agree” with the corresponding empirical means, belongs to [5]. We establish general oracle inequalities and specialize them for different sets of linear functionals. It turns out that universal inequalities of [1] and [4] can be derived from our general oracle inequalities using a specific choice of the set of linear functionals. The results indicate that in the Gaussian white noise model (1) the problem of aggregation of *arbitrary* estimators in $\mathbb{L}_p$, $p \in (2, \infty]$ can be rather difficult. In this case remainder terms in the oracle inequalities depend on the family $\mathcal{F}_\Theta$ and, in general, can be rather large. We prove a lower bound and show that dependence of the remainder terms on $\mathcal{F}_\Theta$ is, in a sense, unavoidable. Thus ”efficient” aggregation of *arbitrary* estimators in $\mathbb{L}_p$, $p \in (2, \infty]$ is impossible. We also show that in the $\mathbb{L}_2$–framework a slight modification of the proposed aggregation procedure satisfies the exact oracle inequality (2) with $C = 1$ and the remainder r_ε that cannot be improved in the minimax sense.

References

[1] Devroye, L. and Lugosi, G. (1996), *A universally acceptable smoothing factor for kernel density estimation.*. Ann. Statist. **24**, 2499–2512.

[2] Devroye, L. and Lugosi, G. (1997), *Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes.* Ann. Statist. **25**, 2626–2637.

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

[4] Hengartner, N. and Wegkamp, M. (2001), *Estimation and selection procedures in regression: an L_1 approach.* Canad. J. Statist. **29**, 621–632.

[5] Nemirovski, A. S. (1985), *Nonparametric estimation of smooth regression functions.* Soviet J. Comput. Systems Sci. **23** , no. 6, 1–11; translated from *Izv. Akad. Nauk SSSR Tekhn. Kibernet.* 1985, no. 3, 50–60, 235 (Russian).

Capturing Functions of Few Variables in High Dimensions

Ronald DeVore

(joint work with Guergana Petrova and Przemyslaw Wojtaszczyk)

The numerical solution of many scientific problems can be reformulated as the approximation of a function f, defined on a domain in $\mathbb{R}^N$, with N large. Most classical numerical methods and the corresponding theory of approximation for N-variate functions deteriorate severely with the growth of N. This is the so-called *curse of dimensionality.* On the other hand, the functions f that arise as solutions to real world problems are thought to be much better behaved than a general N-variate function in the sense that they depend on only a few parameters or variables or they can be well approximated by such functions. This has led to a concerted effort to develop a theory and algorithms which approximate such functions well without suffering the effect of the curse of dimensionality. There are many impressive approaches (see [1, 2, 8, 4, 6, 7, 9] as representative) which are being developed in a variety of settings. One of the main challenges in this approach is to identify the significant variables in a compuationally friendly way. In this talk we shall consider a simple model for the above problem of variable reduction and show that under thsi model it is possible to capture such functions with a modest cost from N.

We shall assume that f is a function defined on $\Omega := [0,1]^N$ but it depends on just ℓ of the coordinate variables: $f(x_1,\dots,x_N) = g(x_{i_1},\dots,x_{i_\ell})$ where $i_1,\dots,i_\ell$ are not known to us. We are given a budget m of questions we can ask about f. Each question takes the form: What is the value of f at a point of our choosing? We want then to approximate f from these point values. We are interested in what are the best questions to ask and to what error can we capture f as we allow the number m of questions to increase. We shall measure the error of approximation in the norm

$$\|\cdot\| := \|\cdot\|_{C(\Omega)}, \tag{1}$$

where $C(\Omega)$ is the space of continuous functions on Ω and the norm is the supremum norm on Ω. Similar questions in other norms are of interest but will not be discussed. The quantitative results we obtain will be made under some smoothness assumption on g in the form of Lipschitz or higher smoothness.

We shall primarily be interested in the case where the m points are assigned in advance. However, we shall also make some remarks on possible gains in reducing the number of function evaluations by proceeding adaptively. A second problem we shall consider is when f is not a function of ℓ variables but it can be approximated to some tolerance ϵ by such a function. We seek again sets of points where the knowledge of the values of f at such points will allow us to approximate f well.

If g is in Lipα and the coordinates $i_1, \dots, i_\ell$ are known to us, then by asking for the values of f at $m = L^\ell$ appropriately spaced points, we could recover f to the accuracy $\|g\|_{\mathrm{Lip}_\alpha} L^{-\alpha}$ in the norm of $C(\Omega)$. We shall show that we can obtain similar estimates even when the coordinates $i_1, \dots, i_\ell$ are not known to us. However, to achieve this performance we shall have to ask slightly more questions. For example, $m = C(\ell) L^{\ell+1} (\log_2 N)^2$ point values will give the accuracy $\|g\|_{\mathrm{Lip}_\alpha} L^{-\alpha}$. The additional factor $L(\log_2 N)^2$ is the price our algorithm pays for not knowing the coordinates $i_1, \dots, i_\ell$.

The construction of the favorable set of points where we ask for the values of f in the general case is based on having a family $\mathcal{A}$ of partitions of $\{1, \dots, N\}$ into ℓ disjoint sets which have certain separation properties on k-tuples of integers. The requirements on $\mathcal{A}$ are closely related to perfect hashing [3, 5]. This approach has been used in Computer Science to identify change variables in discrete settings and the corresponding algorithms are known as JUNTAs.

References

[1] M. Belkin and P. Niyogi, *Laplacian Eigenmaps for Dimensionality Reduction and Data Representation.* Neural Computation **15**(2003),1373–1396.

[2] Coifman, R.R., M. Maggioni, *Diffusion wavelets.* Appl. Comp. Harm. Anal., (21(1)), **2006**, 53-94.

[3] M.Fredman and J.Komlos, *On the size of separating systems and families of perfect hash functions.* SIAM J. Alg. Disc. Meth. **5** (1984), 61–68.

[4] Eitan Greenshtein, *Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint.* Ann. Stat., **34** (2006), 2367–2386.

[5] J.Körner and K.Marton, *New bounds for Perfect Hashing via Information Theory.* Europ. J. Combinatorics **9** (1988), 523–530.

[6] M. Kolountzakis, E. Markakis and A. Mehta, *Learning Symmetric k-juntas in time $n^{o(k)}$*. Proceedings of "Interface between Harmonic Analysis and Number Theory" (2005)

[7] E. Mossel, R. O'Donnell and R. Servedio, *Learning Juntas.* Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC) San-Diego (2003), 206–212.

[8] E. Novak and H. Wozniakowski, *Tractability of Multivariate Problems vol. I: Linear Information.* European Math. Soc. 2008.

[9] Todor, R. A. and C. Schwab, *Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients.* ETH Report, 2007.

Least squares estimation and variable selection under minimax concave penalty

Cun-Hui Zhang

1. Summary. Variable selection and estimation of a sparse high-dimensional signal vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is well understood if it is directly observed with white noise. If most components of $\boldsymbol{\beta}$ are zero, threshold estimators at level $\lambda_{univ} = \sigma\sqrt{2\log p}$ find the exact set of nonzeros with high probability when the minimum absolute value of the nonzero components of $\boldsymbol{\beta}$ is slightly greater than λ_{univ}. If $\boldsymbol{\beta}$ belongs to an ℓ_r ball of radius R, threshold estimators at a certain level λ_{mm} approximately attain the minimax risk. In linear regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n,\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n),$$

with either a deterministic $\boldsymbol{X}$ or a random one independent of $\boldsymbol{\varepsilon}$, these results can be extended up to an $O(1)$ factor with a certain concave penalized estimator, provided that the number of variables to be selected or R^r/λ_{mm}^r is no greater than a certain d_*. Under certain conditions on the design matrix $\boldsymbol{X}$, the order of this d_* could be as high as $n/\log(p/n)$.

2. Minimax concave penalty. Consider penalized loss

$$L_n(\boldsymbol{\beta}, \lambda) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2/(2n) + \sum_{j=1}^p \rho(|\beta_j|; \lambda),$$

with $(\partial/\partial t)\rho(t;\lambda) = \dot\rho(t;\lambda) \ge 0$ for $t > 0$ and $\dot\rho(0+;\lambda) \ge 0$. All critical points of the penalized loss must satisfy the estimating equation

$$\begin{cases} \boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n = \mathrm{sgn}(\beta_j)\dot\rho(|\beta_j|;\lambda), & \beta_j \ne 0 \\ |\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/n| \le \dot\rho(0+;\lambda), & \beta_j = 0 \end{cases} \tag{1}$$

via sub-differentiation. Let $\widehat{\boldsymbol{\beta}}$ be a solution of (1), $\widehat{A} = \{j : \widehat{\beta}_j \ne 0\}$ and $A^o = A^o(\boldsymbol{\beta}) = \{j : \beta_j \ne 0\}$. The residual vector of $\widehat{\boldsymbol{\beta}}$ can be written as

$$\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \mathbb{P}_{\widehat{A}})(\boldsymbol{X}_{\widehat{A}^c}\boldsymbol{\beta}_{\widehat{A}^c} + \boldsymbol{\varepsilon}) + \text{bias}$$

where bias $= \boldsymbol{X}_{\widehat{A}}(\boldsymbol{X}_{\widehat{A}}'\boldsymbol{X}_{\widehat{A}})^{-1}\big(\mathrm{sgn}(\widehat{\beta}_j)\dot\rho(|\widehat{\beta}_j|;\lambda)\big)'$, $\boldsymbol{X}_A = (\boldsymbol{x}_j, j \in A)$ and $\boldsymbol{b}_A = (b_j, j \in A)'$. Since the bias could be correlated with $\boldsymbol{x}_j, j \in \widehat{A}^c$, it may lead to false negative (FN) selection in case of $j \in A^o \setminus \widehat{A}$ and/or false positive (FP) with $\widehat{A} \setminus A^o \ne \emptyset$. The bias may persist and cause significant FN, even in the noiseless case $\sigma = 0$ as shown in Table 1.

If $\dot\rho(t;\lambda) = 0$ for $t > \gamma\lambda$ [7], the bias of penalized estimators can be alleviated if $|\widehat{\beta}_j| \ge \gamma\lambda$ for sufficiently many j. However, such penalty functions are necessarily non-convex.

The minimax concave penalty (MCP),

$$\rho(t;\lambda) = \lambda\int_0^t (1 - x/(\gamma\lambda))_+ dx, \tag{2}$$

TABLE 1. Sparse recovery: $\sigma = 0$, $\lambda = 0+$

$\boldsymbol{y} = \boldsymbol{X\beta}$ with iid $N(0,1)$ entries in $\{\boldsymbol{X}, \beta_j, j \in A^o\}$
FN $= \#\{j : \beta_j \neq 0 = \widehat{\beta}_j\}$, FP allowed for correct recovery

$n, p, \lvert A^o\rvert$ $n/\log(p/n)$	$100, 2000, 15$ 33.38		$100, 2000, 28$ 33.38		$200, 10000, 40$ 51.12	
	mc+	Lasso	mc+	Lasso	mc+	Lasso
$\%\{\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}\}$	**100**	**51**	**73**	0	**100**	0
$\widehat{E}[\mathrm{FN}\lvert\widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}]$		2	19	13		18
$\widehat{E}[\#\mathrm{steps}\lvert\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}]$	32	65	87		102	
$\widehat{E}[\#\mathrm{steps}\lvert\widehat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}]$		144	513	153		311

minimizes the maximum concavity $\max_{0<t<\gamma\lambda}(-1)(\partial/\partial t)^2\rho(t;\lambda)$ among all penalties satisfying $\dot{\rho}(t;\lambda) = 0\ \forall t > \gamma\lambda$ and $\dot{\rho}(0+;\lambda) = \lambda$.

3. Algorithm. Let

$$\lambda^{(x)} \oplus \widehat{\boldsymbol{\beta}}^{(x)},\ \lambda^{(0)} \oplus \widehat{\boldsymbol{\beta}}^{(0)} = \|\boldsymbol{X}'\boldsymbol{y}/n\|_\infty \oplus 0 \tag{3}$$

be a continuous path of solutions of the estimation equation (1). For a given penalty level λ, we define a penalized estimator

$$\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(x_\lambda)}, \quad x_\lambda = \inf\{x : \lambda^{(x)} \leq \lambda\}. \tag{4}$$

Consider a quadratic spline penalty $\rho(t)$ satisfying $(d/dt)\rho(t) > 0\ \forall t > 0$ and $(d/dt)\rho(0+) = 1$. Let $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$. Since $\rho(t)$ is a quadratic spline, the solution path (3) is a linear spline in $\mathbb{R}^{p+1}$. It uniquely exists, ends at a least squares fit at $\lambda = 0+$ and changes its "active set" one at a time almost everywhere. We use a penalized linear unbiased selection (PLUS) algorithm [16] to compute this path. The PLUS algorithm is an extension of the LARS [10, 11, 6] and costs nearly the same number of operations per step (Table 1). We call (4) the PLUS solution of the penalized least squares problem and describe blow properties of such specific solutions. The PLUS solution is a global minimizer of the penalized loss if the penalized loss is convex, but we do not seek global minimization in general. The ℓ_1, MCP, and SCAD [7] are all quadratic splines, with 1, 2 and 3 knots respectively. The PLUS solution for the MCP is called MC+ as in Table 1.

4. Variable selection. We first state a lower bound for selection consistency. Define the minimax average probability of incorrect selection

$$\mathscr{R}(n,p,d,\epsilon) = \inf_{\widehat{A}} \sup_{\beta_* \geq \epsilon} \binom{p}{d}^{-1} \sum_{\|\boldsymbol{\beta}\|_0 = d} P_{\boldsymbol{\beta}}\Big\{\widehat{A} \neq A^o(\boldsymbol{\beta})\Big\}, \tag{5}$$

where $\beta_* = \min_{\beta_j \neq 0}|\beta_j|$, $\|\boldsymbol{\beta}\|_0 = \#\{j : \beta_j \neq 0\}$, $A^o(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$, and the infimum is taken over all Borel maps from $(\boldsymbol{X}, \boldsymbol{y})$ to $\widehat{A} \subseteq \{1, \ldots, p\}$.

Theorem 1. *Suppose $\boldsymbol{X}$ is either deterministic with $\|\boldsymbol{x}_j\|^2 = n$ or random with $E\|\boldsymbol{x}_j\|^2 = n$. Suppose $p - d \to \infty$. Then, $\mathscr{R}(n,p,d,\epsilon) \to 0$ implies $\epsilon \geq (1/2 + o(1))\sigma\sqrt{(2/n)\log(p-d)}$.*

This result [17] is an extension of [15] from standard Gaussian $\boldsymbol{X}$ to general $\boldsymbol{X}$.

We consider the following sufficient conditions for selection consistency:
(S1) For all $j \le p$, $\|\boldsymbol{x}_j\|^2 = n$; for all $|A| \le d^*$ and $\|\boldsymbol{b}_A\| = 1$,

$$0 < c_* \le \|\boldsymbol{X}_A \boldsymbol{b}_A\|^2/n \le c^* < \infty.$$

(S2) For $d_* = 2d^*/\{1 + (1 + I_{\sigma>0})c^*/c_*\}$,

$$\|\boldsymbol{\beta}\|_0 = \#\{j : \beta_j \ne 0\} \le d_*.$$

(S3) For $\lambda_1 = \sigma\sqrt{(2/n)\log(p - \|\boldsymbol{\beta}\|_0)}$ and $\lambda_2 = \sigma\sqrt{(2/n)\log(p/\|\boldsymbol{\beta}\|_0)}$,

$$P\{\lambda_1 \le \widehat{\lambda} \le \lambda_3\} \to 1,\ \lambda_3 \ge \lambda_1 \vee ((2 + \epsilon_0)\sqrt{c^*}\lambda_2),\ \epsilon_0 > 0.$$

(S4) With $\|\boldsymbol{\beta}\|_0 \le a_n \to \infty$,

$$\beta_* \ge \sigma\sqrt{\max\{\mathrm{diag}((\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o}/n)^{-1})\}2\log a_n} + \gamma\lambda_3.$$

(S5) The MCP (2) is used with $\gamma \ge c_*^{-1}\sqrt{4 + c + */c^*}$.

Theorem 2. *Suppose that Conditions S1-S5 hold. Let* $A^o = \{j : \beta_j \ne 0\}$ *and* $\widehat{\boldsymbol{\beta}}^o = (\boldsymbol{X}'_{A^o}\boldsymbol{X}_{A^o})^{-1}\boldsymbol{X}_{A^o}\boldsymbol{y}$ *as an oracle LSE. Let* $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\lambda}+)$ *be the MC+ estimator at the estimated level* $\widehat{\lambda}$, *where* $\widehat{\boldsymbol{\beta}}(\lambda)$ *is as in (4). Then,*

$$P\{\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}),\ \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^o\} \to 1. \tag{6}$$

The interpretation of (6) in the case of $\sigma = 0$ is $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. Condition S1 is called the sparse Riesz condition in [18]. It is a slightly more general version of the restricted isometry condition [3] about $\delta_{d^*} = (c^* - 1) \vee (1 - c_*)$. Condition S2 holds if $\delta_{3d_*} \le 3/7$ for $\sigma > 0$ or $\delta_{2d_*} < 1/2$ for $\sigma = 0$. Condition S3 allows the use of the deterministic $\widehat{\lambda} = \lambda_{univ} = \sigma\sqrt{(2/n)\log p}$ or an upper confidence bound of λ_1. For $\log(\|\boldsymbol{\beta}\|_0) = o(1)\log(p - \|\boldsymbol{\beta}\|_0)$, Condition S4 requires $\boldsymbol{\beta}_* \ge \sqrt{c^*}(2 + o(1))\gamma\lambda_1$ while the lower bound is $\boldsymbol{\beta}_* \ge (1/2 + o(1))\lambda_1$ in Theorem 1; For $\lambda_2 = o(\lambda_1)$ and $\|\boldsymbol{\beta}\|_0 \le p/2$, $\beta_* \ge 2\gamma\lambda_1$ implies Condition S4. Theorem 2 is a sharper version of the selection consistency theorem in [16]. Selection consistency in linear regression has been proved under different sets of conditions in [8, 12, 21, 14] for the Lasso and in [20] for a certain stepwise regression strategy.

5. Estimation. We first state lower bounds for the ℓ_q loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q = \sum_{j=1}^p |\widehat{\beta}_j - \beta_j|^q$ in ℓ_r balls $\Theta_{r,R} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_r \le R\}$.

Theorem 3. *Suppose* $\boldsymbol{X}$ *is deterministic with* $\|\boldsymbol{x}_j\|^2 = n$. *Let* $\lambda_{mm} = \sigma\sqrt{(2/n)\log(\sigma^r p/(n^{r/2}R^r))}$. *Then, for* $r \vee 1 \le q$ *and* $0 < \epsilon < 1$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{r,R}} E_{\boldsymbol{\beta}}\big\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big\|_q^q \ge (1 + o(1))R^r\lambda_{mm}^{q-r}, \tag{7}$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\Theta_{r,R}} P_{\boldsymbol{\beta}}\Big\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \ge (1 - \epsilon)R^r\lambda_{mm}^{q-r}\Big\} \ge \frac{1 - \epsilon + o(1)}{3^q}, \tag{8}$$

as $(n\lambda_{mm}^2/\sigma^2, R^r/\lambda_{mm}) \to (\infty, \infty)$.

The lower bound (7) is an extension of [5] from the case of orthonormal design $\boldsymbol{x}'_k\boldsymbol{x}_k/n = I_{j=k}$.

Consider the following conditions for upper bounds:
(E2) $R^r/\lambda_{mm}^r = |B| \le d_*$ with the d_* in (S2) and a certain $B \subset \{1, \ldots, p\}$.
(E3) $\lambda_0 = 2\sqrt{c^*}\{\lambda_{mm}(1 + \sqrt{2c_*}) + \epsilon_1\sigma/\sqrt{n}\}$ with a small $\epsilon_1 > 0$.
(E4) $\lambda(1 - |t|/(\gamma\lambda))_+ \le \dot{\rho}(|t|; \lambda) \le \lambda$ with $\gamma \ge c_*^{-1}\sqrt{4 + c_*/c^*}$.

Theorem 4. *Suppose that Conditions S1 and E2-E4 hold. Let* $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda)$ *be as in (4). Then, for* $0 < r < q \le 2$,

$$\inf_{|B| \le d_*} \inf_{\boldsymbol{\beta} \in \Theta_{r,R,B}} P\left\{ \begin{array}{c} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_q^q \le M_q R^r \lambda_{mm}^{q-r} \\ \|\widehat{\boldsymbol{\beta}}\|_0 \le |B| d^*/d_* \end{array} \right\} \to 1,$$

as $n\lambda_{mm}^2/\sigma^2 \to \infty$, *where* $\Theta_{r,R,B} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_{B^c}\|_r \le R, \|\boldsymbol{\beta}_{B^c}\|_\infty \le \lambda_{mm}\}$ *and* $M_q = 2^{(q-1)_+}\{M^q(1/2 + c^*/c_*)^{1-q/2} + 1\}$ *with* $M = 3\sqrt{c^*}/c_* + (3\sqrt{2} + 1)\sqrt{c^*/c_*} + o(1)$.

Performance bounds for the Lasso and Dantzig estimators have been obtained in [1, 2, 4, 9, 13, 18, 19] among others for $\lambda \ge \lambda_{univ} = \sigma\sqrt{(2/n)\log p}$.

References

[1] Bickel, P., Ritov, Y. and Tsybakov, A.B. (2007), *Simultaneous analysis of Lasso and Dantzig selector.* Ann. Statist. To appear.

[2] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007), *Sparsity oracle inequalities for the lasso.* Electron. J. Statist. **1**, 169-194 (electronic).

[3] Candes, E. and Tao, T. (2005), *Decoding by linear programming.* IEEE Trans. Info. Theory **51**, 4203 - 4215.

[4] Candes, E. and Tao, T. (2007), *The Dantzig selector: statistical estimation when p is much larger than n* (with discussion). Ann. Statist. **35**, 2313-2404.

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

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), *Least angle regression* (with discussion). Ann. Statist. **32**, 407-499.

[7] Fan, J. and Li, R. (2001), *Variable selection via nonconcave penalized likelihood and its oracle properties.* J. Amer. Statist. Assoc. **96**, 1348-1360.

[8] Meinshausen, N. and Buhlmann, P. (2006), *High dimensional graphs and variable selection with the Lasso.* Ann. Statist. **34**, 1436-1462.

[9] Meinshausen, N. and Yu, B. (2006), *Lasso-type recovery of sparse representations for high-dimensional data.* Ann. Statist., to appear.

[10] Osborne, M., Presnell, B. and Turlach, B. (2000a), *A new approach to variable selection in least squares problems.* IMA Journal of Numerical Analysis **20**, 389-404.

[11] Osborne, M., Presnell, B. and Turlach, B. (2000b), *On the lasso and its dual.* Journal of Computational and Graphical Statistics **9** (2), 319-337.

[12] Tropp, J.A. (2006). *Just relax: convex programming methods for identifying sparse signals in noise.* IEEE Trans. Inform. Theory **52**, 1030-1051.

[13] van de Geer, S. (2008), *High-dimensional generalized linear models and the Lasso.* Ann. Statist. **36**, 614-645.

[14] Wainwright, M. (2006), *Sharp thresholds for high-dimensional and noisy recovery of sparsity.* Technical Report 708, Department of Statistics, University of California, Berkeley.

[15] Wainwright, M.J. (2007), *Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting.* Technical report, Department of Statistics, University of California, Berkeley.

[16] Zhang, C.-H. (2007), *Penalized linear unbiased selection.* Technical Report No. 2007-003. Department of Statistics, Rutgers University.

[17] Zhang, C.-H. (2007), *Information-theoretic optimality of variable selection with concave penalty.* Technical Report No. 2007-008. Department of Statistics, Rutgers University.

[18] Zhang, C.-H. and Huang, J. (2008), *The sparsity and bias of the LASSO selection in high-dimensional regression.* Ann. Statist. **36**, 1567-1594.

[19] Zhang, T. (2007), *Some performance bounds for least squares regression with L1 regularization.* Technical Report 2007-005, Department of Statistics and Biostatistics, Rutgers University.

[20] Zhang, T. (2008), *Adaptive forward-backward greedy algorithm for learning sparse representations.* Technical Report No. 2008-002, Department of Statistics, Rutgers University.

[21] Zhao, P. and Yu, B. (2006), *On model selection consistency of LASSO.* J. Machine Learning Research **7**, 2541-2567.

Reporter: Charles Mitchell

Participants

Prof. Dr. Ery Arias-Castro
Department of Mathematics
University of California, San Diego
9500 Gilman Drive
La Jolla , CA 92093-0112
USA

Prof. Dr. Peter L. Bartlett
Department of Statistics
University of California
# 367 Evans Hall
Berkeley , CA 94720-3860
USA

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

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

Prof. Dr. Florentina Bunea
Department of Statistics
Florida State University
Tallahassee FL 32306-4330
USA

Prof. Dr. Emmanuel Candes
Applied Mathematics 217-50
California Institute of Technology
Pasadena , CA 91125
USA

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

Prof. Dr. Ronald A. DeVore
Department of Mathematics
Texas A & M University
College Station , TX 77843-3368
USA

Prof. Dr. David L. Donoho
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA

Prof. Dr. Lutz Dümbgen
Universität Bern
Institut für Mathematik
Alpeneggstrasse 22
CH-3012 Bern

Prof. Dr. Ursula Gather
Fachbereich Statistik
Technische Universität Dortmund
44221 Dortmund

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

Prof. Dr. Alexander Goldenshluger
Dept. 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

Prof. Dr. Wolfgang Härdle
Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität zu Berlin
Spandauer Str. 1
10178 Berlin

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

Prof. Dr. Anatoli Iouditski
Lab. de Modelisation et Calcul
L J K
Univ. Joseph Fourier Grenoble I
B.P. 53
F-38041 Grenoble Cedex 9

Prof. Dr. Iain M. Johnstone
Department of Statistics
Stanford University
Sequoia Hall
Stanford , CA 94305-4065
USA

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

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

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

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

Karim Lounici
Laboratoire de Probabilites et
Modeles Aleatoires
Universite Paris VII
175 rue du Chevaleret
F-75013 Paris Cedex

Prof. Dr. Gabor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
E-08005 Barcelona

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

Lukas Meier
Departement Mathematik
ETH-Zentrum
Rämistr. 101
CH-8092 Zürich

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

Charles Mitchell
Departement Mathematik
ETH-Zentrum
Rämistr. 101
CH-8092 Zürich

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

Prof. Dr. Partha Niyogi
Dept. of Math. and Comp. Science
The University of Chicago
Ryerson Hall
1100 East 58th St.
Chicago , IL 60637
USA

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

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

Dr. Massimiliano Pontil
Department of Computer Science
University College London
Gower Street
GB-London WC1E 6BT

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

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

Prof. Dr. Philippe Rigollet
Department of Operations Research
and Financial Engineering
Princeton University
Princeton NJ 08544
USA

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

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

Joseph Salmon
Laboratoire de Probabilites et
Modeles Aleatoires
Universite Paris VII
175 rue du Chevaleret
F-75013 Paris Cedex

Dr. Melanie Schienle
Wirtschaftswissenschaftl. Fakultät
Lehrstuhl für Statistik
Humboldt-Universität zu Berlin
Spandauer Str. 1
10178 Berlin

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

Dr. Ingo Steinwart
Los Alamos National Laboratory
Information Sciences Group, CCS-3
Mail Stop B 265
Los Alamos NM 87545
USA

Prof. Dr. Alexandre B. Tsybakov
CREST
Timbre J340
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. Nicolas Vayatis
Centre de Mathématiques et de Leurs
Applications (CMLA)
Ecole Normale Superieure de Cachan
61, Avenue du President Wilson
F-94235 Cachan Cedex

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

Prof. Dr. Marten Wegkamp
Dept. of Statistics
Florida State University
Tallahassee , FL 32306-3033
USA

Prof. Dr. Yuhong Yang
School of Statistics
313 Ford Hall
224 Church Street S.E.
Minneapolis , MN 55455
USA

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

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

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