Mathematisches Forschungsinstitut Oberwolfach

Report No. 12/2018

DOI: 10.4171/OWR/2018/12

Statistical Inference for Structured High-dimensional Models

Organised by
Anatoli Juditsky, Saint Martin d'Hères
Alexandre Tsybakov, Malakoff
Cun-Hui Zhang, Piscataway

11 March – 17 March 2018

Abstract. High-dimensional statistical inference is a newly emerged direction of statistical science in the 21 century. Its importance is due to the increasing dimensionality and complexity of models needed to process and understand the modern real world data. The main idea making possible meaningful inference about such models is to assume suitable lower dimensional underlying structure or low-dimensional approximations, for which the error can be reasonably controlled. Several types of such structures have been recently introduced including sparse high-dimensional regression, sparse and/or low rank matrix models, matrix completion models, dictionary learning, network models (stochastic block model, mixed membership models) and more. The workshop focused on recent developments in structured sequence and regression models, matrix and tensor estimation, robustness, statistical learning in complex settings, network data, and topic models.

Mathematics Subject Classification (2010): 62Gxx (in particular, 62G05, 62G08, 62G10).

Introduction by the Organisers

The workshop *Statistical Inference for Structured High-Dimensional Models*, organized by Anatoli Juditsky (Université Grenoble-Alpes), Alexandre Tsybakov (CREST, ENSAE), and Cun-Hui Zhang (Rutgers University), was held March 11th – March 17th, 2018. The workshop aimed to highlight recent achievements in high-dimensional inference for structured statistical models based on the interplay of techniques from mathematical statistics, optimization theory and high-dimensional probability, and to bring together researchers to exchange the ideas and to explore open mathematical problems. These goals were largely achieved.

The workshop was well attended by 52 participants with broad geographic representation from three continents. Twenty five talks were presented, and seven PhD students shortly presented their work in a "Young researcher's series" on Tuesday evening. The talks can be roughly chategorized into the following topics, which the workshop was focused on.

Estimation and inference in structured sequence and high-dimensional regression models: Pierre Bellec reports recent advances on the noise-barrier and signal bias of the Lasso and other convex estimators; Emmanuel Candes presents an asymptotic theory in logistic regression in the regime where the number of data points is of the same order as the number of unknown parameters; Richard Samworth studies the least square estimation in isotonic regression in general dimensions; Bin Yu studies local identifiability analysis of dictionary learning.

Matrix and tensor estimation: Vladimir Koltchinskii discusses asymptotically efficient estimation of functionals of high-dimensional covariance; Zongming Ma reports recent developments in local asymptotic normality in spiked random matrix models; Vladimir Spokoiny studies large ball probability with applications to inference for spectral projectors; Martin Wahl presents relative perturbation bounds with applications to empirical covariance operators; Dong Xia studies noisy low rank tensor completion; Anru Zhang discusses singular value decomposition for high-dimensional tensor data.

Robust inference: Rina Foygel Barber studies robust inference with the knockoff filter; Olivier Collier presents recent work on sparse functional estimation and robust variance estimation; Arnak Dalalyan studies statistically and computationally efficient estimation of multidimensional linear functionals; Stanislav Minsker considers robust modifications of U-statistics and their applications.

Statistical leaning in complex settings and computational issues: Chao Gao presents convergence rates of variational posteriors; Alexandra Carpentier studies hypothesis testing with Gaussian mixture models; Andrea Montanari studies feasibility in weak recovery of high-dimensional signals; Boaz Nadler develops an asymptotic theory of projection pursuit in high dimensions; Richard Nickl studies information operators and statistical inverse problems; Johannes Schmidt-Hieber presents a statistical theory for deep neural networks; Yihong Wu studies optimal estimation of Gaussian mixtures via denoised method of moments.

Network data, topic models and other applications: Mladen Kolar studies estimation and inference for differential networks; Jing Lei discusses nonparametric network representation and estimation using graph root distribution; Florentina Bunea presents optimal estimation of structured loading matrices with applications to overlapping clustering and topic models; Zheng Ke develops a spectral approach to optimal topic estimation.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons

Foundation for supporting Yihong Wu in the “Simons Visiting Professors” program at the MFO.

Workshop: Statistical Inference for Structured High-dimensional Models

Table of Contents

Abstracts

On statistical and computational complexity of two problems: estimating multidimensional linear functionals and robust estimation of a mean

Arnak Dalalyan
(joint work with Olivier Collier)

Let us consider the problem of estimating a linear functional in the Gaussian sequence model. This means that we observe n noisy vectors $Y_1, \dots, Y_n \in \mathbb{R}^p$ such that

$$Y_i = \theta_i + \sigma\xi_i, \qquad i = 1, \dots, n, \tag{1}$$

where $\theta_i = \mathbf{E}[Y_i]$ are the unknown signals. The noise variables ξ_i are assumed to be iid Gaussian $\mathcal{N}_p(0, I_p)$ and the noise level is assumed to be known. Our goal in this model is to estimate the linear functional $L(\Theta) = \sum_{i\in[n]} \theta_i$ in the scenario of column sparsity of the matrix $\Theta = [\theta_1, \dots, \theta_n] \in \mathbb{R}^{p\times n}$. This means that the number of nonzero columns of Θ, denoted by s, is much smaller than n.

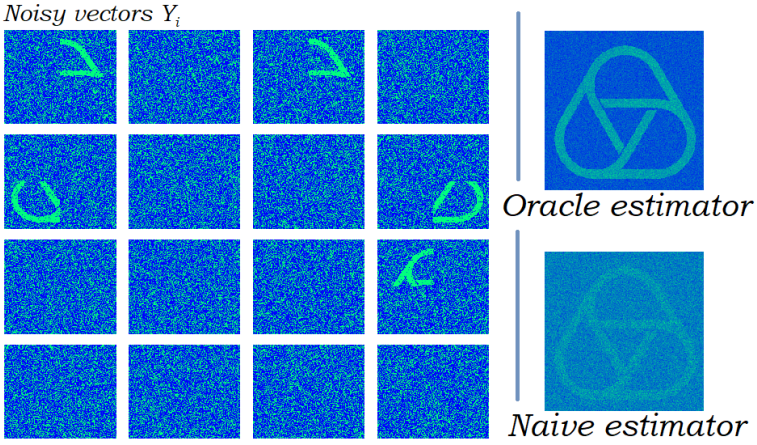

Figure 1. An illustration of the problem of linear functional estimation. In this example $n = 10^3$, $p = 5\times 10^4$ and $s = 50$. We show on the left a subsample of 16 vectors Y_i each of which is obtained by vectorizing the corresponding image and, on the right, the results obtained by the naive estimator $\sum_{i\in[n]} Y_i$ and the oracle estimator $\sum_{i:\theta_i\neq 0} Y_i$.

Our first result shows that the minimax risk defined by

$$r^{\mathrm{mmx}}(n, p, s) = \inf_{\hat{L}} \sup_{\Theta} \mathbf{E}[\|\hat{L} - L(\Theta)\|_2^2]$$

admits the following lower bound: there is a universal constant $c > 0$ such that

$$r^{\mathrm{mmx}}(n, p, s) \geq c\sigma^2(s^2 \log(1 + \tfrac{n}{s^2}) + sp).$$

We then analyze some estimator having polynomial in (s, p, n) computational complexity, such as the block soft and hard thresholding, and show that for the best choice of the tuning parameter they have a worst case risk of order $\sigma^2(s^2\{p\log(1+n^2p/s^4)\}^{1/2}+sp)$. To date, this is the best known upper bound on the minimax risk over all possible computationally efficient estimators. Along with this result, we define a new estimator $\hat{L}^{GSS}$, termed greedy subset selector, for which the following property holds: there is a universal constant C such that for every $\delta\in(0,1)$ and for every s-column sparse matrix Θ,

$$\mathbf{P}\big[\|\hat{L}^{GSS}-L(\Theta)\|_2^2\le C\sigma^2\big(s^2\log(n/\delta)+sp\big)\wedge\big(n\log(1/\delta)+np\big)\big]\ge 1-\delta.$$

Combined with the lower bound mentioned above, we see that the nonasymptotic minimax rate is between $s^2\wedge n+sp$ and $s^2\wedge(np)+sp$, up to possible logarithmic factors. This also shows that the GSS estimator is rate optimal in the regime $s=O(\sqrt{n}\vee p)$. In the regime of the large sparsity, $s\ge const(\sqrt{n}\vee p)$, the rate of the minimax risk remains unknown (we only know that it is between $n+sp$ and $s^2\wedge np$).

We then focus on the problem of robust estimation of the mean of a Gaussian distribution. The model, similar to (1), reads as

$$Y_i=\mu\theta_i+\sigma\xi_i,\qquad i=1,\dots,n, \tag{2}$$

where $\mu\in\mathbb{R}^p$ is the unknwon mean. The goal now is to estimate μ considering that the nuisance parameter Θ is s-column sparse as before. The sparsity s plays here the role of the number of outliers. Without loss of generality, we assume hereafter that $\sigma=1$.

Using the results of [2], one can show that the minimax risk in this problem, defined as

$$r_\mu^{mmx}(n,p,s)=\inf_{\hat\mu}\sup_{\mu,\Theta}\mathbf{E}[\|\hat\mu-\mu\|_2^2]$$

is of the order (up to possible log terms)

$$r_\mu^{mmx}(n,p,s)\overset{\log}{\asymp}\frac{p}{n}+\Big(\frac{s}{n}\Big)^2,$$

and that this rate can be attained by Tukey's median. It is also known that the coordinate-wise median or the geometric median have a much larger risk, of the order

$$\frac{p}{n}+p\Big(\frac{s}{n}\Big)^2.$$

The most important shortcoming of Tukey's median is that its computational complexity is prohibitively large even for moderate values of p. To improve on this, we continue our efforts [5, 1] in using convex programming based approaches to solve the robust estimation problem. Our main result is the introduction of a new estimator, that can be seen as an iterative group-soft thresholding, that satisfies the following property: for every $\nu\in(0,1)$ (close to zero), if we perform $K=\log_2(1/\nu)+\log\log p$ iterations of our algorithm IGST, the resulting estimator

$\hat{\mu}^{\mathrm{IGST}}$ satisfies

$$\mathbf{E}[\|\hat{\mu}^{\mathrm{IGST}} - \mu\|_2^2] \lesssim \frac{p}{n} + \left(\frac{s}{n}\right)^2 + \left(\frac{s^4 p}{n^4}\right)^{1-\nu}.$$

All these results are discussed and proved in [3], while analogous problems in the Poisson model are studied by [4].

References

[1] S. Balmand, A. Dalalyan, *Convex programming approach to robust estimation of a multivariate Gaussian model*, arXiv:1512.04734, 2015.

[2] M. Chen, C. Gao and Z. Ren, *Robust Covariance and Scatter Matrix Estimation under Huber's Contamination Model*, arXiv:1506.00691, 2015.

[3] O. Collier, A. Dalalyan, *Rate-optimal estimation of p-dimensional linear functionals in a sparse Gaussian model*, arXiv:1712.05495, 2017.

[4] O. Collier, A. Dalalyan, *Estimating linear functionals of a sparse family of Poisson means*, Stat Inference Stoch Process (2018), `https://doi.org/10.1007/s11203-018-9173-0`.

[5] A. Dalalyan, Y. Chen, *Fused sparsity and robust estimation for linear models with unknown variance*, Advances in Neural Information Processing Systems 25 (NIPS 2012).

Sparse functional estimation and robust variance estimation

Olivier Collier

(joint work with Laëtitia Comminges and Alexandre B. Tsybakov)

Adaptive estimation in the sparse mean model and in sparse regression exhibits some interesting effects. We consider estimation of a sparse vector, of its l_2-norm and of the noise variance in the sparse linear model. We establish the optimal rates of adaptive estimation when adaptation is considered with respect to the noise level, the noise distribution and sparsity. These rates turn out to be different from the minimax non-adaptive rates when they are known. A crucial issue is the ignorance of the noise level. Moreover, knowing or not knowing the noise distribution can also influence the rate. For example, the rates of estimation of the noise level can differ depending on whether the noise is Gaussian or sub-Gaussian without a precise knowledge of the distribution. We also show that in the problem of estimation of a sparse vector under the l_2-risk when the variance of the noise in unknown, the optimal rate depends dramatically on the design.

Robust modifications of U-statistics and their applications

Stanislav Minsker

(joint work with Xiaohan Wei)

Let Y be a d-dimensional random vector with unknown mean μ and covariance matrix Σ. Results presented in this talk are motivated by the task of designing an estimator of Σ that admits tight deviation bounds in the operator norm $\|\cdot\|$ under minimal assumptions on the underlying distribution, such as existence of only 4th moments of the coordinates of Y. To address this task, we propose

"robust" versions of the operator-valued U-statistics, obtain non-asymptotic guarantees for their performance, and demonstrate implications of these general results for structural covariance estimation problems.

Consider a sequence of i.i.d. random variables $X_1, \dots, X_n$ ($n \geq 2$) taking values in a measurable space $(\mathcal{S}, \mathcal{B})$. Assume that $H : \mathcal{S}^m \to \mathbb{H}^d$ (where $2 \leq m \leq n$ and $\mathbb{H}^d$ is the set of all $d \times d$ self-adjoint matrices) is a $\mathcal{S}^m$-measurable permutation symmetric kernel, meaning that $H(x_1, \dots, x_m) = H(x_{\pi_1}, \dots, x_{\pi_m})$ for any $(x_1, \dots, x_m) \in \mathcal{S}^m$ and any permutation π. Let $\Psi(x) = \begin{cases} \frac{x^2}{2} - \frac{|x|^3}{6}, & |x| \leq 1, \\ \frac{1}{3} + \frac{1}{2}(|x| - 1), & |x| > 1 \end{cases}$ be the analogue of Huber's loss function. For any $A \in \mathbb{H}^d$ with the spectral decomposition $A = \sum_j \lambda_j u_j u_j^*$, we define $\Psi(A) = \sum_j \Psi(\lambda_j) u_j u_j^*$. Given $\theta > 0$, let

$$\widehat{U}_{n,\theta} = \mathrm{argmin}_{U \in \mathbb{H}^d} \, \mathrm{trace} \left[\sum_{(i_1, \dots, i_m) \in I_n^m} \Psi\Big(\theta \left(H(X_{i_1}, \dots, X_{i_m}) - U \right) \Big) \right]$$

be the "robust" version of the U-statistic

$$U_n = \frac{(n-m)!}{n!} \sum_{(i_1, \dots, i_m) \in I_n^m} H(X_{i_1}, \dots, X_{i_m}).$$

The following result holds: let $k = \lfloor n/m \rfloor$, and assume that $t > 0$ satisfies $\frac{dt}{k} \leq \frac{1}{104}$. Then for any $\sigma \geq \left\| \mathbb{E} \left(H(X_1, \dots, X_m) - \mathbb{E} H(X_1, \dots, X_m) \right)^2 \right\|^{1/2}$ and $\theta_* := \frac{1}{\sigma} \sqrt{\frac{2t}{k}}$, $\left\| \widehat{U}_{n,\theta_*} - \mathbb{E} H \right\| \leq 23 \sigma \sqrt{\frac{t}{k}}$ with probability at least $1 - (4d+1)e^{-t}$. Several extensions and statistical applications of this result are discussed.

References

[1] S. Minsker, X. Wei, *Robust Modifications of U-statistics and Applications to Covariance Estimation Problems*, arXiv:1801.05565 (2018).

[2] S. Minsker, *Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries*, arXiv:1605.07129 (2017).

What do we *really* know about logistic regression?

Emmanuel J. Candès

(joint work with Pragya Sur)

Every student in statistics or data science learns early on that when the sample size n largely exceeds the number p of variables, fitting a logistic model produces estimates that are approximately unbiased. Every student also learns that there are formulas to predict the variability of these estimates which are used for the purpose of statistical inference; for instance, to produce p-values for testing the significance of regression coefficients. Although these formulas come from large sample asymptotics, in which the number n of observations is increasingly large and the number p of variables under study is held constant, we are often told that

we are on reasonably safe grounds when n is large in such a way that $n \geq 5p$ or $n \geq 10p$. This paper shows that this is far from the case, and consequently, inferences routinely produced by common software packages are often unreliable. This is a very siginificant problem since applied researchers everywhere routinely fit high-dimensional models.

Consider a logistic model with independent features in which n and p become increasingly large in a fixed ratio. Then we show that

(1) the MLE is biased,
(2) the variability of the MLE is far greater than classically predicted,
(3) and the commonly used likelihood-ratio test (LRT) is not distributed as a chi-square.

The bias of the MLE is extremely problematic as it yields completely wrong predictions for the probability of a case based on observed values of the covariates.

In this talk, we present a new theory, which asymptotically predicts (1) the bias of the MLE, (2) the variability of the MLE, and (3) the distribution of the LRT. We empirically also demonstrate that these predictions are extremely accurate in finite samples. Further, an appealing feature is that these novel predictions depend on the unknown sequence of regression coefficients only through a single scalar, the overall strength of the signal. This suggests very concrete procedures to adjust inference; we describe one such procedure learning a single parameter from data and producing accurate inference.

Our theory is presented in [1], see also [2] and we give below one salient result. Imagine we have n independent observations $(y_i, \boldsymbol{X}_i)$ where $y_i \in \{0, 1\}$ is the response variable and $\boldsymbol{X}_i \in \mathbb{R}^p$ the vector of predictor variables. The logistic model posits that the probability of a case conditional on the covariates is given by

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{X}_i) = \sigma(\boldsymbol{X}_i'\boldsymbol{\beta}),$$

where $\sigma(t) = e^t/(1 + e^t)$ is the standard sigmoidal function. We describe the asymptotic properties of the MLE and the LRT in a high-dimensional regime, where n and p both go to infinity in such a way that $p/n \to \kappa$. We work with independent observations $\{\boldsymbol{X}_i, y_i\}$ from a logistic model such that $\mathbb{P}(y_i = 1 \mid \boldsymbol{X}_i) = \rho'(\boldsymbol{X}_i'\boldsymbol{\beta})$. We assume here that $\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is the p-dimensional identity matrix. (This means that the columns of the matrix $\boldsymbol{X}$ of covariates are unit-normed in the limit of large samples.). The exact scaling of $\boldsymbol{X}_i$ is not important: the important scaling is the overall strength of the signal and we assume that the p regression coefficients (recall that p increases with n) are scaled in such a way that

$$\lim_{n\to\infty} \mathrm{Var}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \gamma^2, \tag{1}$$

where γ is fixed. It is useful to think of the parameter γ as the signal strength. Another way to express (1) is to say that $\lim_{n\to\infty} \|\boldsymbol{\beta}\|^2/n = \gamma^2$.

Theorem 1. *Assume the dimensionality and signal strength parameters κ and γ are such that $\gamma < g_{MLE}(\kappa)$ (the region where the MLE exists asymptotically).*

Assume the logistic model described above where the empirical distribution of $\{\beta_j\}$ *converges weakly to a distribution* Π *with finite second moment. Suppose further that the second moment converges in the sense that as* $n \to \infty$, $\mathrm{Ave}_j(\beta_j^2) \to \mathbb{E}\beta^2$, $\beta \sim \Pi$. *Then for any pseudo-Lipschitz function* ψ *of order* 2,[1] *the marginal distributions of the MLE coordinates obey*

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{\beta}_j - \alpha_\star \beta_j, \beta_j) \xrightarrow{\text{a.s.}} \mathbb{E}[\psi(\sigma_\star Z, \beta)], \quad Z \sim \mathcal{N}(0,1), \tag{2}$$

where $\beta \sim \Pi$, *independent of* Z.

Above, α and $\sigma_\star$ are solutions to a simple system of nonlinear equations—naturally, the system depends on κ and γ—and can be easily computed.

In a nutshell, this new theory establishes that in some sense, $\hat{\beta}_j$ is asymptotically normal with mean $\alpha_\star$ and standard deviation $\sigma_\star$. This is not at all what classical predicts. More rigorously,

- this result quantifies the exact bias of the MLE in some statistical sense. This can be seen by taking $\psi(t,u) = t$ in (2), which leads to
$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star \beta_j) \xrightarrow{\text{a.s.}} 0,$$
and says that $\hat{\beta}_j$ is centered about $\alpha_\star\, \beta_j$.
- Second, our result also provides the asymptotic variance of the MLE marginals after they are properly centered. This can be seen by taking $\psi(t,u) = t^2$, which leads to
$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star \beta_j)^2 \xrightarrow{\text{a.s.}} \sigma_\star^2.$$
- Third, our result establishes that upon centering the MLE around $\alpha_\star \boldsymbol{\beta}$, it becomes decorrelated from the signal $\boldsymbol{\beta}$. This can be seen by taking $\psi(t,u) = tu$, which leads to
$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \alpha_\star \beta_j)\, \beta_j \xrightarrow{\text{a.s.}} 0.$$

References

[1] Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *arXiv preprint arXiv:1803.06964*, 2018.

[2] Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *arXiv preprint arXiv:1706.01191*, 2017.

[1] A function $\psi : \mathbb{R}^m \to \mathbb{R}$ is said to be pseudo-Lipschitz of order k if there exists a constant $L > 0$ such that for all $\boldsymbol{t}_0, \boldsymbol{t}_1 \in \mathbb{R}^m$, $\|\psi(\boldsymbol{t}_0) - \psi(\boldsymbol{t}_1)\| \le L\left(1 + \|\boldsymbol{t}_0\|^{k-1} + \|\boldsymbol{t}_1\|^{k-1}\right)\|\boldsymbol{t}_0 - \boldsymbol{t}_1\|$.

Statistical theory for deep neural networks with ReLU activation function

Johannes Schmidt-Hieber

Large databases and increasing computational power have recently resulted in astonishing performances of deep neural networks (DNNs) for a broad range of learning tasks, including image and text classification, speech recognition and game playing. Available theoretical results on neural networks cannot explain these successes. A mathematical theory is, however, essential for several purposes. Firstly, one wants to explain and understand several effects that are empirically observed in trained networks. Secondly, a lot of expert knowledge is required to set up the architecture of a network. Theoretical results should in the long run replace this expert knowledge and lead to automatic recommendations of preferable initial network configurations. Thirdly, one would like to have mathematical tools that compare deep networks to other existing techniques. Finally, a mathematical theory should be able to detect scenarios for which neural networks are not optimal and point to possible limitations. The mathematical understanding should then give insight on how to further improve the performance of neural networks. All four aspects of a mathematical theory for multilayer neural networks are currently unsolved.

A neural network requires the choice of an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and the network architecture. The main interest is in understanding the rectifier linear unit (ReLU) $\sigma(x) = \max(x, 0)$ because there is a clear gain in deep networks using the ReLU instead of sigmoidal activation function, cf. [2]. In computer science, multilayer neural networks are defined as directed acyclic graphs. For theoretical purposes, it is, however, more convenient to work with the following equivalent algebraic definition. For $\mathbf{v} = (v_1, \ldots, v_r) \in \mathbb{R}^r$, define the shifted activation function $\sigma_{\mathbf{v}} : \mathbb{R}^r \to \mathbb{R}^r$ as $\sigma_{\mathbf{v}}(\mathbf{y}) = (\sigma(y_i - v_i))_{i=1,\ldots,r}$. The network architecture $(L, \mathbf{p})$ consists of a positive integer L called the *number of hidden layers* or *depth* and a *width vector* $\mathbf{p} = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$. A neural network with network architecture $(L, \mathbf{p})$ is then any function of the form

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad \mathbf{x} \mapsto f(\mathbf{x}) = W_{L+1}\sigma_{\mathbf{v}_L} W_L \sigma_{\mathbf{v}_{L-1}} \cdots W_2 \sigma_{\mathbf{v}_1} W_1 \mathbf{x}, \tag{1}$$

where W_i is a $p_i \times p_{i-1}$ weight matrix and $\mathbf{v}_i \in \mathbb{R}^{p_i}$ is a shift vector. Network functions are therefore build by alternating matrix-vector multiplications with the action of the non-linear activation function σ. A network with one hidden layer is called shallow and the term multilayer refers to $L > 1$.

The key problem is to fit a neural network given training data. This means that the network architecture is chosen beforehand and the weight matrices and shift vectors that constitute the set of network parameters are learned/estimated from the sample. In the supervised learning setting n independent and identically distributed copies of pairs $(\mathbf{X}, Y)$ are observed. Here, $\mathbf{X}$ is a d-dimensional random vector modelling the input of the network and Y denotes the corresponding output. The output can be a real number or a class label. The statistical challenge is to reconstruct/learn the regression function $f(\mathbf{x}) = E[Y|\mathbf{x} = \mathbf{x}]$ from the sample.

In practice, the loss induced by the log-likelihood is minimized using stochastic gradient descent (SGD). Because of the non-convex function space, these gradient descent methods converge to one of the many local minima. It is now widely believed that the risk of most of the local minima is not much larger than the risk of the global minimum, cf. [1]. For the theory it is convenient to ignore SGD and to study the empirical risk minimiser instead.

To develop theory, we consider nonparametric regression. If the errors are Gaussian, the log-likelihood induced loss is the least squares loss. Given a network architecture $(L, \mathbf{p})$ and a number of non-zero/active parameters s, the $(L, \mathbf{p}, s)$-neural network estimator is defined as the minimizer of the empirical least squares risk over all networks with architecture $(L, \mathbf{p})$, s non-zero network parameters and all non-zero network parameters bounded in absolute value by a constant.

In principle, we could study the convergence rate of the neural network estimator assuming that the true regression function is β-smooth (e.g. in the Hölder sense). Since in practical applications of DNNs the input dimension is large any statistical method will suffer from the curse of dimensionality. Convergence rates are hence extremely slow. We argue that it is therefore irrelevant whether a method attains these rate. Instead we study a class for which faster rates are achievable. Many objects for which DNNs give state of the art results have a modular or hierarchical structure. To build text, for instance, we first can generate lines, from lines letters, from letters words, from words sentences and from sentences paragraphs. The key observation is that only few objects are combined to build an object on a higher abstraction level. To build a word for instance, few letters are combined. Mathematically speaking, the structural assumption on the regression function is that it has a decomposition of the form $f = g_q \circ \ldots \circ g_0$ for functions $g_i = (g_{ij})_j : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$, where each function g_{ij} has Hölder smoothness β_i and is t_i-variate.

If $\beta_i^* := \beta_i \prod_{\ell=i+1}^{q} (\beta_\ell \wedge 1)$ we can show that the prediction risk of the $(L, \mathbf{p}, s)$-neural network estimator achieves the minimax rate $\max_i n^{-2\beta_i^*/(2\beta_i^*+t_i)}$ up to logarithmic factors provided that the network depth L is proportional to $\log n$, the width of the network is large (sufficiently large powers of n) and $s \asymp \max_i n^{t_i/(2\beta_i^*+t_i)} \log n$. The precise statement can be found in [3]. The convergence rate only depends on the "effective dimension" t_i but not on d_i. This can lead to fast rates that are not affected by the curse of dimensionality. As special cases, one can show that DNNs achieve (near) optimal rates for common structural constraints such as (generalized) additive models.

Networks of the form (1) with ReLU activation function generate piecewise linear functions. Fitting piecewise linear functions does typically not give optimal rates if the true function is more than twice differentiable. Because of the nonlinearity, DNNs achieve (near) optimal rates for all smoothness indices. For that we need the network depth to be proportional to $\log n$.

References

[1] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, Y. LeCunn *The loss surface of multilayer networks*, Aistats **38** (2015), 192–204.
[2] X. Glorot, A. Bordes, Y. Bengio, *Deep sparse rectifier neural networks*, Aistats **15** (2011), 315–323.
[3] J. Schmidt-Hieber, *Nonparametric regression using deep neural networks with ReLU activation function*, arXiv preprint 1708.06633, (2017).

The noise barrier and the large signal bias of the Lasso and other convex estimators

Pierre C. Bellec

Convex estimators such as the Lasso, the matrix Lasso and the group Lasso have been studied extensively in the last two decades, demonstrating great success in both theory and practice. This paper introduces two quantities, the noise barrier and the large scale bias, that provides novel insights on the performance of these convex regularized estimators.

In sparse linear regression, it is now well understood that the Lasso achieves fast prediction rates, provided that the correlations of the design satisfy some Restricted Eigenvalue or Compatibility condition, and provided that the tuning parameter is at least larger than some universal threshold. Using the two quantities introduced in the paper, we show that the compatibility condition on the design matrix is actually unavoidable to achieve fast prediction rates with the Lasso. In other words, the ℓ_1-regularized Lasso must incur a loss due to the correlations of the design matrix, measured in terms of the compatibility constant. This results holds for any design matrix, any active subset of covariates, and any positive tuning parameter.

It is now well known that the Lasso enjoys a dimension reduction property: if the target vector is s-sparse, the prediction rate of the Lasso with tuning parameter λ is of order $\lambda\sqrt{s}$, even if the ambient dimension p is much larger than p. Such results require that the tuning parameters is greater than some universal threshold. We characterize sharp phase transitions for the tuning parameter of the Lasso around a critical threshold dependent on s. If λ is equal or larger than this critical threshold, the Lasso is minimax over s-sparse target vectors. If λ is equal or smaller than critical threshold, the Lasso incurs a loss of order $\sigma\sqrt{s}$ –which corresponds to a model of size s– even if the target vector is more sparse than s.

Remarkably, the lower bounds obtained in the paper also apply to random, data-driven tuning parameters. Additionally, the results extend to convex penalties beyond the Lasso.

Robust inference with the knockoff filter

Rina Foygel Barber
(joint work with Emmanuel Candès, Richard Samworth)

We consider the variable selection problem, which seeks to identify important variables influencing a response Y out of many candidate features $X_1, \ldots, X_p$. We wish to do so while offering finite-sample guarantees about the fraction of false positives—selected variables X_j that in fact have no effect on Y after the other features are known. More formally, after observing the data, we select a set $\widehat{S} \subset \{1, \ldots, p\}$ indexing the features that we believe to be directly associated with Y, and would like to control the false discovery rate,

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \{j : \text{ feature } X_j \text{ is null}\}|}{\max\{1, |\widehat{S}|\}}\right],$$

where a feature X_j is said to be *null* if $X_j \perp\!\!\!\perp Y \mid X_{-j}$, i.e. X_j is independent from the response Y after controlling for the $p-1$ remaining features, $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$.

When the number of features p is large (perhaps even larger than the sample size n), and we have no prior knowledge regarding the type of dependence between Y and X, the model-X knockoffs framework [2] nonetheless allows us to select a model with a guaranteed bound on the false discovery rate, as long as the distribution of the feature vector $X = (X_1, \ldots, X_p)$ is exactly known. This model selection procedure operates by constructing "knockoff copies" of each of the p features, denoted by $\widetilde{X}_1, \ldots, \widetilde{X}_p$, which are then used as a control group to ensure that the model selection algorithm is not choosing too many irrelevant features. The method operates by drawing the p-dimensional knockoff feature vector $\widetilde{X}$, conditionally on the observed features X, in such a way that the joint distribution of $(X, \widetilde{X})$ is *pairwise exchangeable*, i.e. for each j,

$$(X, \widetilde{X}) \stackrel{\mathrm{d}}{=} (X_1, \ldots, X_{j-1}, \widetilde{X}_j, X_{j+1}, \ldots, X_p, \widetilde{X}_1, \ldots, \widetilde{X}_{j-1}, X_j, \widetilde{X}_{j+1}, \ldots, \widetilde{X}_p),$$

meaning that by swapping the jth feature X_j with its knockoff copy $\widetilde{X}_j$, the overall distribution is unchanged. The combined $2p$ many features (p original features and p knockoffs) along with the response Y are then fed as input to a model selection algorithm, and the algorithm's ability to filter out the knockoffs from its selected set is used to indicate its likely FDR.

In this work, we study the practical setting where the distribution of X could only be estimated, rather than known exactly, and the knockoff copies of the X_j's are therefore constructed somewhat incorrectly. Our results, which are free of any modeling assumption whatsoever, show that the resulting model selection procedure incurs an inflation of the false discovery rate that is proportional to our errors in estimating the distribution of each feature X_j conditional on the remaining features $\{X_k : k \neq j\}$. The model-X knockoffs framework is therefore robust to errors in the underlying assumptions on the distribution of X, making it an effective method for many practical applications, such as genome-wide association

studies, where the underlying distribution on the features $X_1, ..., X_p$ is estimated accurately but not known exactly.

More concretely, suppose that P_X is the true distribution of feature vector X, and let Q_X be an estimate of this distribution. Suppose that our knockoff features $\widetilde{X}$ are constructed such that the pairwise exchangeability property is satisfied relative to the *estimated* distribution Q_X of the features X (since the true distribution P_X is not known). In other words, we choose a conditional distribution $\widetilde{P}_{\widetilde{X}|X}$ such that the joint distribution $Q_X \times \widetilde{P}_{\widetilde{X}|X}$ satisfies pairwise exchangeability, i.e.

$$\text{If } X \sim Q_X \text{ and } \widetilde{X} \mid X \sim \widetilde{P}_{\widetilde{X}|X}, \text{ then pairwise exchangeability is satisfied.}$$

Assuming that the true distribution P_X is well approximated by Q_X, then the true joint distribution of $(X, \widetilde{X})$, which is given by $P_X \times \widetilde{P}_{\widetilde{X}|X}$, must approximately satisfy pairwise exchangeability. To quantify this, we define the following measure of divergence: for each feature j, let

$$\widehat{\mathrm{KL}}_j = \sum_{i=1}^{n} \log\left(\frac{P_j(X_{ij} \mid X_{i,-j}) Q_j(\widetilde{X}_{ij} \mid X_{i,-j})}{Q_j(X_{ij} \mid X_{i,-j}) P_j(\widetilde{X}_{ij} \mid X_{i,-j})}\right),$$

where P_j and Q_j are the jth conditionals of P_X and Q_X, respectively—that is, the true and estimated conditional distribution of X_j given X_{-j}, and the indices $i = 1, \ldots, n$ denote the n i.i.d. data points. In other words, this divergence measures whether swapping X_j with $\widetilde{X}_j$ (across all n observed data points) would substantially change the likelihood of the data. The notation $\widehat{\mathrm{KL}}_j$ suggests the KL divergence, and indeed, $\mathbb{E}[\widehat{\mathrm{KL}}_j]$ is equal to the KL divergence between the distribution of $(X_{i,*}, \widetilde{X}_{i,*})_{i=1,\ldots,n}$, and the distribution of the same random vectors with X_j and $\widetilde{X}_j$ swapped. Of course, it's clear that this divergence is zero if pairwise exchangeability for the jth feature is satisfied exactly.

Our main results prove that $\widehat{\mathrm{KL}}_j$ exactly characterizes the performance of the knockoff filter. First, as an upper bound, we show that false discoveries are controlled, at least among the set of features X_j whose knockoffs are constructed to be nearly pairwise exchangeable in the sense that $\widehat{\mathrm{KL}}_j$ is small:

Theorem 1.

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \{j : \textit{ feature } X_j \textit{ is null, and } \widehat{\mathrm{KL}}_j \le \epsilon\}|}{\max\{1, |\widehat{S}|\}}\right] \le \alpha \cdot e^{\epsilon},$$

where α is the target FDR level, and $\epsilon \ge 0$ is arbitrary.

Our proof is based on a novel leave-one-out technique, inspired by leave-one-out analyses of the Benjamini-Hochberg method [4] for multiple testing, e.g. the analysis of [5].

Second, we show a lower bound that (up to constant factors) matches this upper bound, proving that if a worst-case algorithm can lose FDR control whenever $\max_{\text{null } j} \widehat{\mathrm{KL}}_j$ is likely to be $\ge \epsilon$:

Theorem 2. *If* $\mathbb{P}\{\widehat{\mathrm{KL}}_j \geq \epsilon\} \geq c > 0$ *for some null* j*, then there exists a model selection procedure that controls FDR at level* α *when* $X \sim Q_X$ *is the true feature distribution, but if* P_X *is the true feature distribution, has FDR* $\geq \alpha \cdot (1 + c(1 - e^{-\epsilon}))$*.*

Many open questions remain for this problem—in particular, in future work, we hope to address the question of whether the divergence measure $\widehat{\mathrm{KL}}_j$ is too pessimistic in practical settings, where we might be likely to use model selection algorithms that are severely constrained in how they use the available data (e.g. selecting features based only on coarse summary statistics), for which our worst-case lower bound result (Theorem 2) above will no longer apply. In practical settings, are there less conservative measures of our approximation error in the model for the feature vector X, that will lead to FDR control results once we assume some constraints on the model selection algorithm?

References

[1] R. F. Barber, E. J. Candès, and R. J. Samworth, *Robust inference with knockoffs*, `arXiv:1801.03896`.

[2] E. J. Candès, Y. Fan, L. Janson, and J. Lv. *Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection*, To appear in Journal of the Royal Statistical Society: Series B, 2018.

[3] R.F. Barber and E. J. Candès. *Controlling the false discovery rate via knockoffs*, Annals of Statistics **43** (2015), 2055–2085.

[4] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: Series B **57** (1995): 289–300.

[5] J. A. Ferreira and A. H. Zwinderman, *On the Benjamini–Hochberg method*, Annals of Statistics **34** (2006), 1827–1849.

A spectral approach to topic modeling

Zheng Tracy Ke

In the probabilistic topic models, the quantity of interest—a low-rank matrix consisting of topic vectors—is hidden in the text corpus matrix, masked by noise, and Singular Value Decomposition (SVD) is a potentially useful tool for learning such a low-rank matrix. However, the connection between this low-rank matrix and the singular vectors of the text corpus matrix are usually complicated and hard to spell out, so how to use SVD for learning topic models faces challenges.

We overcome the challenge by revealing a surprising insight: there is a low-dimensional *simplex* structure which can be viewed as a bridge between the low-rank matrix of interest and the SVD of the text corpus matrix, and which allows us to conveniently reconstruct the former using the latter. Such an insight motivates a new SVD-based approach to learning topic models.

For asymptotic analysis, we show that under the popular probabilistic model (Hofmann 1999), the convergence rate of the ℓ^1-error of our method matches that of the minimax lower bound, up to a multi-logarithmic term. In showing these results, we have derived new element-wise bounds on the singular vectors and

several large-deviation bounds for weakly dependent multinomial data. Our results on the convergence rate and asymptotical minimaxity are new.

We have applied our method to two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), with encouraging results. In particular, there is a clear simplex structure associated with the SVD of the data matrices, which largely validates our discovery.

Optimal estimation of structured loading matrices with applications to overlapping clustering and topic models

Florentina Bunea

(joint work with Mike Bing, Yang Ning, Marten Wegkamp)

This work introduces a novel estimation method, called LOVE, of the entries and structure of a loading matrix A in a sparse latent factor model $X = AZ + E$, for an observable random vector X in R^p, with correlated unobservable factors $Z \in R^K$, with K unknown, and independent noise E. Each row of A is scaled and sparse. In order to identify the loading matrix A, we require the existence of pure variables, which are components of X that are associated, via A, with one and only one latent factor. Despite the fact that the number of factors K, the number of the pure variables, and their location are all unknown, we only require a mild condition on the covariance matrix of Z, and a minimum of only two pure variables per latent factor to show that A is uniquely defined, up to signed permutations. Our proofs for model identifiability are constructive, and lead to our novel estimation method of the number of factors and of the set of pure variables, from a sample of size n of observations on X. This is the first step of our LOVE algorithm, which is optimization-free, and has low computational complexity of order p^2. The second step of LOVE is an easily implementable linear program that estimates A. We prove that the resulting estimator is minimax rate optimal up to logarithmic factors in p.

The model structure is motivated by the problem of overlapping variable clustering, ubiquitous in data science. We define the population level clusters as groups of those components of X that are associated, via the sparse matrix A, with the same unobservable latent factor, and multi-factor association is allowed. Clusters are respectively anchored by the pure variables, and form overlapping sub-groups of the p-dimensional random vector X. The **L**atent model approach to **OVE**rlapping clustering is reflected in the name of our algorithm, LOVE.

The third step of LOVE estimates the clusters from the support of the columns of the estimated A. We further guarantee cluster recovery with zero false positive proportion, and with false negative proportion control. The practical relevance of LOVE is illustrated through the analysis of an RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.

We also consider the related, but different, problem of estimation in topic models. If one observes n independent multinomials of dimension p, the topic models postulate a certain factorization of the expectation of the $p \times n$ data matrix. We

provide conditions under which the factors are identifiable, and concentrate on the estimation of one of them, known as the word-topic matrix. We adapt LOVE to this problem, and provide a computationally efficient algorithm that yields minimax adaptive estimators of the word-topic matrix.

References

[1] M. Bing, F. Bunea, Y. Ning, M. Wegkamp, *Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering*, https://arxiv.org/abs/1704.06977 (2018).
[2] M. Bing, F. Bunea, M. Wegkamp, *A new approach to optimal estimation in topic models with unknown number of topics*, Preprint (2018).

Two problems in weak recovery

Andrea Montanari

We study the problem of reconstructing an unknown high-dimensional parameters vector "better than random," in two different contexts.

(1) Generalized linear measurements [with M. Mondelli]
Unknown object: $\theta_0 \in \mathbb{R}^d$, $\|\theta_0\|_2 = \sqrt{d}$. $(y_i, x_i)_{i\le n}, x_i \sim N(0, I_d/d)$, $y_i \sim P(\cdot|\langle\theta_0, x_i\rangle)$.
(2) Group-synch on grids [with E. Abbe, L. Masoulie, A. Sly, N. Srivastava]
Unknown object $(\theta_i)_{i\in V}$, $\theta_i \in \mathcal{G} \subseteq \mathbb{R}^{m\times m}$ a compact group; measurements: Y_{ij} with $\mathbb{E}(Y_{ij}|\theta) = c\theta_i^{-1}\theta_j$ for $G = (V, E)$ a known graph.
We obtain several results when $G = \mathbb{Z}^d$ is the d-dimensional grid.

Convergence Rates of Variational Posterior Distributions

Chao Gao

We study convergence rates of variational posterior distributions for nonparametric and high-dimensional inference. We formulate general conditions on prior, likelihood, and variational class that characterize the convergence rates. Under similar "prior mass and testing" conditions considered in the literature, the rate is found to be the sum of two terms. The first term stands for the convergence rate of the true posterior distribution, and the second term is contributed by the variational approximation error. For a class of priors that admit the structure of a mixture of product measures, we propose a novel prior mass condition, under which the variational approximation error of the generalized mean-field class is dominated by convergence rate of the true posterior. We demonstrate the applicability of our general results for various models, prior distributions and variational classes by deriving convergence rates of the corresponding variational posteriors.

"Large ball probability" and inference for spectral projectors

Vladimir Spokoiny

Let $X_1, \dots, X_n$ be i.i.d. sample in $\mathbb{R}^p$ with zero mean and the covariance matrix Σ. We consider the problem of confidence estimation of the projectors on a eigenspace of Σ. This paper offers two procedures: one is based on the resampling technique; the other one is Bayesian and uses Bayesian calculus from the conjugated Wishart prior. Accuracy of both methods is evaluated with sharp error bounds. The study heavily uses recent results on "large ball probability" for Gaussian measures in a Hilbert space.

Local identifiability analysis of dictionary learning

Bin Yu

At the workshop, Bin Yu from UC Berkeley gave a talk on "local identifiability analysis of dictionary learning." She started by citing Jerzy Neyman, the founding father of Berkeley Statistics, that the experimental sciences are sources of theoretical problems. Using a functional genomics project to map a cell's destiny based on spatial gene expression images from embryonic fruitflies, she motivated the study of the local idenfiability problem of dictionary learning.

In particular, theoretical properties of learning a dictionary via ℓ_1-minimization are studied. N data points or images are assumed i.i.d. random linear combinations of the K columns from a complete (i.e., square and invertible) reference dictionary, a $K \times K$ matrix D. Here, the random linear coefficients are generated from either the s-sparse Gaussian model or the Bernoulli-Gaussian model. First, for the population case, a sufficient and almost necessary condition is established for the reference dictionary D to be locally identifiable, i.e., a local minimum of the expected ℓ_1-norm objective function. Our condition covers both sparse and dense cases of the random linear coefficients and significantly improves the sufficient condition by Gribonval and Schnass (2010). In addition, it is shown that for a complete μ-coherent reference dictionary, i.e., a dictionary with absolute pairwise column inner-product at most $\mu \in [0, 1)$, local identifiability holds even when the random linear coefficient vector has up to $O(\mu^{-2})$ nonzeros on average. Moreover, the local identifiability results also translate to the finite sample case with high probability provided that the number of signals N scales as $O(K \log K)$.

One and two sided composite-composite tests in Gaussian mixture models

Alexandra Carpentier

(joint work with Nicolas Verzelen, Etienne Roquain, Sylvain Delattre)

Finding an efficient test for a testing problem is often linked to the problem of estimating a given function of the data. When this function is not smooth, it is necessary to approximate it cleverly in order to build good tests.

In this talk, we will discuss two specific testing problems in Gaussian mixtures models. In both, the aim is to test the proportion of null means. The aforementioned link between sharp approximation rates of non-smooth objects and minimax testing rates is particularly well illustrated by these problems.
Consider a distribution ν with support in $\mathbb{R}$. We observe n i.i.d. data of distribution

$$X_i \sim \nu * \mathcal{N}(0,1).$$

Let $\rho > 0$ and $p \in [0,1)$. We consider the two following testing problems :

$$\mathbf{TP} : H_{0,p} : \nu = (1-p)\delta_0 + p\nu' \text{ vs } H_{1,\rho} : \mathrm{Supp}(\nu) \subset [-\rho,\rho]^c,$$

(two sided problem) and

$$\mathbf{TP}+ : H^+_{0,p} : \nu = (1-p)\delta_0 + p\nu', \mathrm{Supp}(\nu') \subset \mathbb{R}^+ \text{ vs } H^+_{1,\rho} : \mathrm{Supp}(\nu) \subset [\rho,+\infty),$$

(one sided problem). These two testing problems correspond to testing a given proportion of null means in the means of the signal, with or without a positivity constraint.
Our objective is then to find the minimax optimal order of ρ such that a non-trivial test exists, i.e. find

$$\rho^*_p = \inf\{\rho > 0 : \exists T \text{ test with }, \inf_{\nu \in H_{0,p}} \mathbb{E}_\nu[T] + \inf_{\nu \in H_{1,\rho}} \mathbb{E}_\nu[1-T] < 1/2\Big\},$$

and

$$\rho^{*,+}_p = \inf\{\rho > 0 : \exists T \text{ test with }, \inf_{\nu \in H^+_{0,p}} \mathbb{E}_\nu[T] + \inf_{\nu \in H^+_{1,\rho}} \mathbb{E}_\nu[1-T] < 1/2\Big\}.$$

When $p = 0$, the problem is akin to simple signal detection [3] - the null hypothesis is simple and we have respectively

$$\rho^*_0 \approx n^{-1/4} \text{ and } \rho^{*,+}_0 \approx n^{-1/2}.$$

We consider rather the specific case $p = 1/2$, which corresponds to a large null hypothesis, and our objectives are (i) to understand how the size of the null hypothesis will affect the rate and (ii) to understand how the shape constraint will impact the testing problems, i.e. how different TP and TP+ are. In this case, solving the testing problems is related to estimating a one or two sided indicator function. The main problem is then on the rate at which these indicator functions can be approximated. It is possible to prove that this task can be performed optimally using Chebychev polynoms (one sided case) and symmetrized Chebychev polynoms (two sided case) - leading therefore to upper and lower bounds on ρ as

$$\rho^*_{1/2} \approx \log(n)^{-1/2} \text{ and } \rho^{*,+}_{1/2} \approx \log(n)^{-3/2}.$$

Very related results can be found in [5, 4]. In this problem, it is interesting to see how clearly the upper and lower bounds can be derived by solving a given functional approximation problem.
This testing problem has consequences for the problems of sparsity testing [1], but also estimation in the Gaussian contamination model, and FDR estimation in multiple testing [2].

References

[1] A. Carpentier, and N. Verzelen, *Adaptive estimation of the sparsity in the Gaussian vector model*, arXiv preprint arXiv:1703.00167 (2017).

[2] A. Carpentier, S. Delattre, E. Roquain, and N. Verzelen, *Adaptive Estimation of a mean in the presence of one-sided Contaminations and application to multiple testing*, working paper (2018).

[3] Y. Ingster, I.A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*, Springer Science & Business Media **169** (2012).

[4] J. Jin, *Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators*, J. R. Statist. Soc. B **70** (2008), part 3 pp. 461-493.

[5] A. Juditsky, and A. Nemirovski, *On nonparametric tests of positivity/monotonicity/convexity*, Annals of statistics (2002), pp. 498-527.

Isotonic regression in general dimensions

Richard J. Samworth

(joint work with Qiyang Han, Tengyao Wang and Sabyasachi Chatterjee)

We study the least squares regression function estimator over the class of real-valued functions on $[0,1]^d$ that are increasing in each coordinate. For uniformly bounded signals and with a fixed, cubic lattice design, we establish that the estimator achieves the minimax rate of order $n^{-\min\{2/(d+2),1/d\}}$ in the empirical L_2 loss, up to poly-logarithmic factors. Further, we prove a sharp oracle inequality, which reveals in particular that when the true regression function is piecewise constant on k hyperrectangles, the least squares estimator enjoys a faster, adaptive rate of convergence of $(k/n)^{\min(1,2/d)}$, again up to poly-logarithmic factors. Previous results are confined to the case $d \leq 2$. Finally, we establish corresponding bounds (which are new even in the case $d = 2$) in the more challenging random design setting. There are two surprising features of these results: first, they demonstrate that it is possible for a global empirical risk minimisation procedure to be rate optimal up to poly-logarithmic factors even when the corresponding entropy integral for the function class diverges rapidly; second, they indicate that the adaptation rate for shape-constrained estimators can be strictly worse than the parametric rate.

This talk is based on work in the paper [1].

References

[1] Q. Han, T. Wang, S. Chatterjee and R. J. Samworth (2017), *Isotonic regression in general dimensions*, `https://arxiv.org/abs/1708.09468`.

Information operators and statistical inverse problems

Richard Nickl

We consider two statistical inverse problems, and discuss the characterisation of the 'information operator' driving the LAN expansion in these models by PDE methods.

First (see [5]), we consider the statistical inverse problem of recovering a function $f : M \to \mathbb{R}$, where M is a smooth compact Riemannian manifold with boundary, from measurements of general X-ray transforms $I_a(f)$ of f, corrupted by additive Gaussian noise. For M equal to the unit disk with 'flat' geometry and $a = 0$ this reduces to the standard Radon transform, but our general setting allows for anisotropic media M and can further model local 'attenuation' effects – both highly relevant in practical imaging problems such as SPECT tomography. We study a nonparametric Bayesian inference method based on standard Gaussian process priors for f. The posterior reconstruction of f corresponds to a Tikhonov regulariser with a reproducing kernel Hilbert space norm penalty that does not require the calculation of the singular value decomposition of the forward operator I_a. We prove Bernstein-von Mises theorems for a large family of one-dimensional linear functionals of f, and they entail that posterior-based inferences such as credible sets are valid and optimal from a frequentist point of view. In particular we derive the asymptotic distribution of smooth linear functionals of the Tikhonov regulariser, which attains the semi-parametric information lower bound. The proofs rely on an invertibility result for the 'Fisher information' operator $I_a^* I_a$ between suitable function spaces, a result of independent interest that relies on techniques from microlocal analysis.

Second (see [4]), the inverse problem of determining the potential $f > 0$ in the partial differential equation

$$\frac{\Delta}{2}u - fu = 0 \text{ on } \mathcal{O} \text{ s.t. } u = g \text{ on } \partial\mathcal{O},$$

where $\mathcal{O}$ is a bounded C^∞-domain in $\mathbb{R}^d$ and $g > 0$ is a given function prescribing boundary values, is considered. The data consist of the solution u corrupted by additive Gaussian noise. A nonparametric Bayesian prior for the function f is devised and a Bernstein - von Mises theorem is proved which entails that the posterior distribution given the observations is approximated by an infinite-dimensional Gaussian measure that has a 'minimal' covariance structure in an information-theoretic sense. The function space in which this approximation holds true is shown to carry the finest topology permitted for such a result to be possible. As a consequence the posterior distribution performs valid and optimal frequentist statistical inference on f in the small noise limit.

For background on nonparametric statistics and Bernstein-von Mises theorems, we refer to [1, 2, 3].

References

[1] I. Castillo, R. Nickl, *Nonparametric Bernstein-von Mises theorems in Gaussian white noise*, Annals of Statistics **41** (2013).

[2] I. Castillo, R. Nickl, *On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures*, Annals of Statistics **42** (2014).

[3] E. Giné, R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*, *Cambridge University Press, New York*, 2016.

[4] R. Nickl, *Bernstein-von Mises theorems for statistical inverse problems I: Schroedinger equation*, (2017), https://arxiv.org/abs/1707.01764

[5] F. Monard, R. Nickl, G.P. Paternain, *Efficient nonparametric Bayesian inference for X-ray transforms*, (2017), https://arxiv.org/abs/1708.06332

Asymptotically efficient estimation of functionals of high-dimensional covariance

Vladimir Koltchinskii

We consider a problem of estimation of functionals of the form $\langle f(\Sigma), B\rangle$ of unknown covariance operator Σ in $\mathbb{R}^d$ based on i.i.d. observations $X_1, \dots, X_n$ sampled from normal distribution with zero mean and covariance Σ. Assuming that f is a smooth function in real line and B is an operator with nuclear norm bounded by a constant, the goal is to develop an asymptotically efficient estimator of the functional $\langle f(\Sigma), B\rangle$ with $\sqrt{n}$ convergence rate in the setting when the dimension d of the space is allowed to grow with n. We achieve this goal by developing a new bias reduction method in high-dimensional problems and constructing an estimator of the form $\langle h(\hat{\Sigma}), B\rangle$ with bias of the order $o(n^{-1/2})$, where $\hat{\Sigma}$ is the sample covariance based on observations $X_1, \dots, X_n$ and h is a sufficiently smooth approximate solution of the equation $\mathbb{E}_\Sigma h(\hat{\Sigma}) = f(\Sigma), \Sigma \in \mathcal{C}_+^d$ on the cone $\mathcal{C}_+^d$ of covariance operators in $\mathbb{R}^d$. This estimator coincides with the usual plug-in estimator $\langle f(\hat{\Sigma}), B\rangle$, when $d = o(n^{1/2})$, but, in the case when $d \geq n^{1/2}$, the bias of the plug-in estimator becomes larger than $n^{-1/2}$ and a non-trivial bias correction is necessary to develop efficient estimators with $\sqrt{n}$ convergence rate. We show that if $d \leq n^\alpha$ for some $\alpha \in (0,1)$ and f belongs to Besov space $B^s_{\infty,1}(\mathbb{R})$ for some $s > \frac{1}{1-\alpha}$, then asymptotically efficient estimation is possible. More details could be found in [1].

References

[1] V. Koltchinskii, *Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators*, *arXiv:1710.09072*.

Asymptotic normality of log-likelihood ratios in spiked random matrix models

Zongming Ma
(joint work with Debapratim Banerjee)

We study likelihood ratio statistics in a number of spiked random matrix models, including Gaussian mixtures and spiked covariance matrix models. We work directly with multi-spiked cases and allow flexible sub-Gaussian priors on the signal component. We derive asymptotic normality for the log-likelihood ratios when the signal-to-noise ratios are below certain thresholds.

Singular value decomposition for high-dimensional high-order data

Anru Zhang

High-dimensional high-order data arise in many modern scientific applications including genomics, brain imaging, and social science. In this talk, we propose a general framework of tensor singular value decomposition (tensor SVD), which aims to extract the hidden low-rank structure from high-dimensional high-order data. To be specific, a low-rank tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is observed with entry-wise corruptions as $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$. Here $\mathbf{Z}$ is the p_1-by-p_2-by-p_3 noisy tensor with $\{Z_{ijk}\}_{i,j,k=1}^{p_1,p_2,p_3} \overset{iid}{\sim} N(0, \sigma^2)$; $\mathbf{X}$ is a fixed tensor with low Tucker ranks in the sense that all fibers of $\mathbf{X}$ along three directions (i.e., counterpart of matrix columns and rows for tensors) lie in low-dimensional subspaces, say U_1, U_2, and U_3, respectively. Our goal is to estimate U_1, U_2, U_3, and $\mathbf{X}$ from the noisy observation $\mathbf{Y}$.

First, we develop comprehensive results on both the statistical and computational limits for tensor SVD [1]. Let the Tucker rank of $\mathbf{X}$ be (r_1, r_2, r_3). The statistical and computational barriers of tensor SVD problem rely on a key factor λ, i.e., the smallest non-zero singular values of matricizations of $\mathbf{X}$, which essentially measures the signal strength of the problem. When $p = \min\{p_1, p_2, p_3\}$, $p_k \le Cp$, $r_k \le Cp^{1/2}$ for $k = 1, 2, 3$ and a constant $C > 0$, our main results can be summarized into the following three phases according to signal-to-noise ratio (SNR): λ/σ.

(1) When $\lambda/\sigma = p^\alpha$ for $\alpha \ge 3/4$, the scenario is referred to as the **strong SNR case**. The fast higher-order orthogonal iteration (HOOI) recovers U_1, U_2, U_3, and $\mathbf{X}$ with the minimax optimal rate of convergence over a general class of low-rank tensors.
(2) When $\lambda/\sigma = p^\alpha$ for $\alpha < 1/2$, we refer to this case as the **weak SNR case**, and propose the minimax lower bound to show that there are no consistent estimators of U_1, U_2, U_3, or $\mathbf{X}$;
(3) When $\lambda/\sigma = p^\alpha$ for $1/2 \le \alpha < 3/4$, the scenario is referred to as the **moderate SNR case**. We provide a computational lower bound to show that no polynomial time algorithm can recover U_1, U_2, U_3 consistently based on an assumption of hypergraphic planted clique detection. Meanwhile, the maximum likelihood estimator, although being computational intractable,

achieves optimal rates of convergence over a general class of low-rank tensors.

Second, we consider the sparse tensor singular value decomposition which allows more robust estimation under sparsity structural assumptions [2]. A procedure named *Sparse Tensor Alternating Truncation for Singular Value Decomposition* (STAT-SVD) was proposed for sparse tensor SVD. The method consists of two steps: (i) a thresholded spectral initialization and (ii) an iterative alternating updating scheme. One crucial part of the procedure is a novel *double projection & thresholding* scheme, which provides a sharp criterion for thresholding in each iteration. Since each step of STAT-SVD only involves basic matrix and tensor operations, such as matricization, multiplication, matrix SVD, and thresholding, the proposed procedure can be implemented efficiently. We study both the theoretical and numerical properties of the proposed procedure. In particular, we prove by an upper bound argument that the proposed procedure can recover the low-rank structures accurately. A lower bound is further developed to show that the proposed estimator is rate optimal for a general class of simultaneously sparse and low-rank tensors.

It is also noteworthy that the results can be further generalized to fourth or higher order tensors, or when the noise is sub-Gaussian distributed.

References

[1] A. Zhang and D. Xia *Tensor SVD: statistical and computational limits*, IEEE Transactions on Information Theory, to appear, 2018.

[2] A. Zhang and R. Han *Optimal Sparse Singular Value Decomposition for High-dimensional High-order Data*, preprint, 2017.

Network representation using graph root distributions

Jing Lei

Consider a random symmetric two-way binary array

$$\mathbf{A} = (A_{ij} : \ i \geq 1, \ j \geq 1),$$

with convention $A_{ii} = 0$. Each upper-diagonal entry of $\mathbf{A}$ is a Bernoulli random variable. The row-column joint exchangeability means that

$$(A_{ij} : \ i \geq 1, \ j \geq 1) \stackrel{d}{=} (A_{\sigma(i)\sigma(j)} : \ i \geq 1, \ j \geq 1)$$

for all finite index permutation mapping σ.

Analogous to the de Finetti theorem, the Aldous-Hoover theorem says that any symmetric exchangeable binary array $\mathbf{A}$ can be generated by sampling independent $(s_i : i \geq 1)$ from $\mathrm{Unif}(0,1)$ (the uniform distribution on $[0,1]$), and sampling A_{ij} independently from a Bernoulli distribution with probability $W(s_i, s_j)$ for a symmetric measurable function $W(\cdot,\cdot) : [0,1]^2 \mapsto [0,1]$. For any given realization of $\mathbf{A}$, we can simply treat W as a non-random parameter.

Two functions W_1 and W_2 generate the same distribution of exchangeable arrays if and only if there exist two measure-preserving mappings h_1, h_2 such that

$$W_1(h_1(s), h_1(s')) = W_2(h_2(s), h_2(s'))\,,\ \text{a.e.}\,.$$

When this is the case, we say W_1 and W_2 are *weakly isomorphic*, denoted as

$$W_1 \overset{w.i.}{=} W_2\,.$$

The notion "$\overset{w.i.}{=}$" defines an equivalence relation on the space of all symmetric functions that map $[0,1]^2$ to $[0,1]$. When W_1 and W_2 are not weakly isomorphic, then they lead to different distributions of exchangeable random graphs. In this case, the sub-graph counts have different distributions under W_1 and W_2. Such a sampling distribution difference can be linked to the cut-distance between two graphons, defined as

$$\begin{aligned}&\delta_\square(W_1, W_2)\\ &= \inf_{h_1,h_2} \sup_{S\times S':\subseteq[0,1]^2} \left| \int_{S\times S'} [W_1(h_1(s), h_1(s')) - W_2(h_2(s), h_2(s'))]\, dsds' \right|\,,\end{aligned}$$

where h_1, h_2 range over all measure-preserving mappings.

We develop an alternative characterization of exchangeable random graphs.

Definition 1. *A Kreĭn space $\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_-$ is the direct sum of two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$. For each (x, y), $(x', y') \in \mathcal{K}$ with $x, x' \in \mathcal{H}_+$ and $y, y' \in \mathcal{H}_-$, the Kreĭn inner product is*

$$\langle (x, y), (x', y') \rangle_{\mathcal{K}} = \langle x, x' \rangle_{\mathcal{H}_+} - \langle y, y' \rangle_{\mathcal{H}_-}\,. \tag{1}$$

The space $\mathcal{K}$ is also a linear normed space isomorphic to $\mathcal{H}_+ \oplus \mathcal{H}_-$ equipped with norm

$$\|(x, y)\|_{\mathcal{K}} = (\|x\|^2_{\mathcal{H}_+} + \|y\|^2_{\mathcal{H}_-})^{1/2}\,.$$

Definition 2. *We call a probability measure F on $\mathcal{K}$ a* graph root distribution *if for two independent samples Z_1 and Z_2 from F*

$$\mathbb{P}(\langle Z_1, Z_2 \rangle_{\mathcal{K}} \in [0, 1]) = 1\,.$$

Definition 3. *We say two distributions F_1, F_2 on $\mathcal{K}$ are equivalent up to orthogonal transform, written as $F_1 \overset{o.t.}{=} F_2$, if there exist orthogonal transforms Q_+ on $\mathcal{H}_+$ and Q_- on $\mathcal{H}_-$, such that $(X, Y) \sim F_1 \Rightarrow (Q_+X, Q_-Y) \sim F_2$.*

We summarize our representation results in the following corollary.

Theorem 1. *There exists a one-to-one correspondence between trace-class graphons (under the equivalence relation "$\overset{w.i.}{=}$") and square-integrable GRD's with uncorrelated positive and negative components (under the equivalence relation "$\overset{o.t.}{=}$").*

Next we show that the Wasserstein distance between two equivalence classes of GRD's induce a stronger topology than the cut distance in the space of graphon equivalence classes. We consider the *orthogonal Wasserstein distance*

$$d_{\mathrm{ow}}(F_1, F_2) := \inf_{\nu \in \mathcal{V}(F_1,F_2)} \inf_{Q_+,Q_-} \mathbb{E}_{(Z_1,Z_2)\sim\nu} \|Z_1 - (Q_+ \oplus Q_-)Z_2\| \,,$$

where Q_+, Q_- range over all orthogonal transforms on $\mathcal{H}_+$, $\mathcal{H}_-$, respectively, and $(Q_+ \oplus Q_-)$ denotes the orthogonal transform as the direct sum of Q_+ and Q_-: $(Q_+ \oplus Q_-)(x, y) = (Q_+x, Q_-y)$.

Theorem 2. *Let F_1, F be two square-integrable GRD's on $\mathcal{K}$, with corresponding graphons W_1, W. Then*

$$\delta_\square(W_1, W) \le (\mathbb{E}_{F_1}\|Z\| + \mathbb{E}_F\|Z\|) d_{\mathrm{ow}}(F_1, F) \,.$$

As a consequence, if $(F_N : N \ge 1)$ are square-integrable GRD's on $\mathcal{K}$ with corresponding graphons $(W_N : N \ge 1)$, then

$$d_{\mathrm{ow}}(F_N, F) \to 0 \;\Rightarrow\; \delta_\square(W_N, W) \to 0 \,.$$

We then show that GRD's can be easily estimated. Given $n \ge 1$, suppose we have observed an $n \times n$ block of $\mathbf{A}$: $\mathbf{A}_n = (A_{ij} : 1 \le i, j \le n)$, where $\mathbf{A}$ is generated from a GRD F. Assume

(A1) For all $j, j' \ge 1$, $\mathbb{E}_{(X,Y)\sim F}(X_j X_{j'}) = \lambda_j \mathbf{1}(j = j')$, $\mathbb{E}_{(X,Y)\sim F}(Y_j Y_{j'}) = \gamma_j \mathbf{1}(j = j')$, $\mathbb{E}_{(X,Y)\sim F}(X_j Y_{j'}) = 0$.

(A2) There exist positive numbers $c_1 \le c_2$, $1 < \alpha \le \beta$ such that

$$c_1 j^{-\alpha} \le (\lambda_j \wedge \gamma_j) \le (\lambda_j \vee \gamma_j) \le c_2 j^{-\alpha} \,,\; (\lambda_j - \lambda_{j+1}) \wedge (\gamma_j - \gamma_{j+1}) \ge c_1 j^{-\beta} \,,\; \forall\, j \ge 1 \,.$$

(A3) $\mathbb{E}_{Z\sim F}\|Z\|^4 < \infty$.

Let $(\hat\lambda_j, \hat a_j)$ be the positive eigenvalue-eigenvector pairs of $\mathbf{A}_n$, and $(\hat\gamma_j, \hat b_j)$ be the negative absolute eigenvalue-eigenvector pairs.

Let $\hat X_i = (\sqrt{\hat\lambda_1}\hat a_{1i}, ..., \sqrt{\hat\lambda_p}\hat a_{pi}, 0, ..., 0)$, $\hat Y_i = (\sqrt{\hat\gamma_1}\hat b_{1i}, ..., \sqrt{\hat\gamma_p}\hat b_{pi}, 0, ..., 0)$. Let $\hat F$ be the probability measure with $1/n$ mass at $(\hat X_i; \hat Y_i)$ for $1 \le i \le n$.

Theorem 3. *Under assumptions (A1-A3), we have, when $p = o\left(n^{\frac{1}{2\beta+\alpha}}\right)$,*

$$d_{\mathrm{o.w.}}(\hat F, F) = O_P(p^{-(\alpha-1)/2} + pn^{-1/(p\vee 3)}) \,.$$

The right hand side is $o_P(1)$ if $p = O(\log n / \log\log n)$.

A sparsity parameter can easily be incorporated the graph root sampling scheme. Let F be a GRD. For a node sample size n and sparsity parameter $\rho_n \in (0, 1)$, the corresponding sparse graph root sampling scheme is essentially generating node sample points from a scaled distribution:

$$\mathbf{A}_{n,i,j} \sim \mathrm{Bernoulli}(\langle \rho_n^{1/2} Z_i, \rho_n^{1/2} Z_j \rangle_{\mathcal{K}}) \,,$$

where $Z_i \overset{iid}{\sim} F$.

Theorem 4. *Under assumptions (A1-A3) with $\beta \geq 3\alpha/2$, if $\rho_n \geq c \log n/n$ for a positive constant c and*

$$p = o\left[n^{1/(2\beta+\alpha)} \wedge (n\rho_n)^{1/(2\beta)}\right]$$

then $d_{\mathrm{w}}(\rho_n^{-1/2}\hat{F}, F) = O_P(p^{\beta-(\alpha-1)/2}(n\rho_n)^{-1/2} + p^{-(\alpha-1)/2} + pn^{-1/(p\vee 3)})$.

References

[1] J. Lei, *Network representation using graph root distributions*, arXiv:1802.09684.

Estimation and inference for differential networks

Mladen Kolar

We present a recent line of work on estimating differential networks and conducting statistical inference about parameters in a high-dimensional setting. First, we consider a Gaussian setting and show how to directly learn the difference between the group structures. A debiasing procedure is presented for construction of an asymptotically normal estimator of the difference. Next, building on the first part, we show how to learn the difference between two graphical models with latent variables. Linear convergence rate is established for an alternating gradient descent procedure with correct initialization. Finally, we discuss how to do statistical inference on the differential networks when data are not Gaussian.

Optimal estimation of Gaussian mixtures via denoised method of moments

Yihong Wu
(joint work with Pengkun Yang)

The Method of Moments [5] is one of the most widely used methods in statistics for parameter estimation, obtained by solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or non-uniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analysis of method of moments are mainly confined to asymptotic normality style of results established under strong assumptions [2, 3].

In this talk I will present some recent results for estimating Gaussians location mixtures with known or unknown variance. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space before executing the method of moments. Not only does this regularization ensures existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in Wasserstein distance via polynomial interpolation and marjorization, we establish the statistical guarantees and

optimality of the proposed procedure. In particular, we prove the following: Given n independent samples drawn from a k-component Gaussian mixture $\pi * N(0, \sigma^2)$, where $\pi \triangleq \sum_{i=1}^{k} w_i \delta_{\mu_k}$ is the latent discrete distribution, the number of components k is a constant, π has a bounded support and σ is bounded,

- if σ is known, then there exists an estimator $\hat{\pi}$ such that

$$\mathbb{E}W_1(\pi, \hat{\pi}) \leq O(n^{-\frac{1}{4k-2}})$$

- if σ unknown, then there exists an estimator $(\hat{\pi}, \hat{\sigma})$, due to Lindsay [6], such that

$$\mathbb{E}W_1(\pi, \hat{\pi}) \leq O(n^{-\frac{1}{4k}}) \qquad \mathbb{E}|\sigma - \hat{\sigma}| \leq O(n^{-\frac{1}{2k}}).$$

Both estimators can be computed in $O(n)$ time. Both rates are minimax optimal, with the lower bound in the known-σ case previously shown in [4]. Furthermore, in the known-σ case, the estimator is automatically *adaptive* to the clustering structure, in the sense that if the k components fall into k_0 well-separated (separated by a constant) clusters [1, 4], then the same procedure fulfills $\mathbb{E}W_1(\pi, \hat{\pi}) \leq O(n^{-\frac{1}{4(k-k_0)+2}})$, which, in the fully separated scenario of $k_0 = k$, reduces to the parametric rate $n^{-\frac{1}{2}}$.

These results can also be viewed as provable algorithms for Generalized Method of Moments [3] which involves non-convex optimization and lacks theoretical guarantees. Extensions to multiple dimensions will be discussed. In particular, we show that for d-dimensional location normal mixtures with identity covariance matrix and with bounded means and $d = O(n)$, it is possible to achieve the worst-case rate

$$O\left(\left(\frac{d}{n}\right)^{\frac{1}{4}} + \left(\frac{1}{n}\right)^{\frac{1}{4k-2}} \log^2(n)\right)$$

which is minimax optimal up to logarithmic factors.

References

[1] Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, pages 221–233, 1995.

[2] Aad W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, Cambridge, United Kingdom, 2000.

[3] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

[4] Philippe Heinrich and Jonas Kahn. Optimal rates for finite mixture estimation. *arXiv:1507.04313*, 2015.

[5] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[6] Bruce G Lindsay. Moment matrices: applications in mixtures. *The Annals of Statistics*, pages 722–740, 1989.

Projection pursuit in high dimensions

Boaz Nadler
(joint work with Gil Kur, Peter J. Bickel)

Projection pursuit is a classical exploratory data analysis method to detect interesting low dimensional structure in multivariate data. Originally, projection pursuit was applied mostly to data of moderately low dimension. Motivated by contemporary applications, we here study its properties in high dimensional settings. Specifically, we analyze the asymptotic properties of projection pursuit on structure-less multivariate Gaussian data with an identity covariance, as both dimension p and sample size n tend to infinity, with $p/n \to \gamma \in [0, \infty]$. Our main results are as follows:

(i) if $\gamma = \infty$, then there exist projections whose corresponding empirical cumulative distribution function can approximate *any* arbitrary distribution;
(ii) if $\gamma \in (0, \infty)$, not all limiting distributions are possible. Yet, depending on the value of γ various non-Gaussian distributions may still be approximated.
(iii) If $\gamma \in (0, \infty)$ but we restrict to sparse projections, involving only few of the p variables, then asymptotically all empirical cdf's are Gaussian.
(iv) if $\gamma = 0$, then asymptotically all projections are Gaussian.

We conjecture that for $\gamma < 1$, all distributions that one may converge to must be a mixture of a Gaussian $N(0, q)$ distribution, where $q < (1 + \sqrt{\gamma})^2$, with possibly a small non-Gaussian component.

Some of our results extend to mean centered sub-Gaussian data and to projections into k dimensions. Hence, in the "small n, large p" setting, unless sparsity is enforced and regardless of the chosen projection index, projection pursuit may detect apparent structure that has no statistical significance. Furthermore, our work reveals fundamental limitations on the ability to detect non-Gaussian signals in high dimensional data, in particular via independent component analysis (ICA) and related non-Gaussian component analysis.

References

[1] J.H. Friedman, J. Tukey, *A projection pursuit algorithm for exploratory data analysis*, IEEE Transactions Computer, **23** (1974), 881–889.
[2] P. Bickel, G. Kur and B. Nadler, *Projection pursuit in high dimensions*, preprint (2018).

Relative perturbation bounds with applications to empirical covariance operators

Martin Wahl
(joint work with Moritz Jirak)

Let Σ be a self-adjoint, positive trace class operator on a separable Hilbert space $\mathcal{H}$. By the spectral theorem, there exists a sequence $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ of positive eigenvalues (which is either finite or infinite and summable), together

with an orthonormal system of eigenvectors $u_1, u_2, \dots$ such that Σ has spectral representation $\Sigma = \sum_{j\geq 1} \lambda_j u_j \otimes u_j$. For $u, v \in \mathcal{H}$, we denote by $u \otimes v$ the rank-one operator defined by $(u \otimes v)x = \langle v, x\rangle u$, $x \in \mathcal{H}$. We suppose that the eigenvectors of Σ form an orthonormal basis of $\mathcal{H}$.

Let $\hat{\Sigma}$ be another self-adjoint, positive trace class operator on $\mathcal{H}$. We consider $\hat{\Sigma}$ as a perturbed version of Σ and write $E = \hat{\Sigma} - \Sigma$ for the additive perturbation. Again, by the spectral theorem, there exists a sequence $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq 0$ of eigenvalues, together with an orthonormal basis of eigenvectors $\hat{u}_1, \hat{u}_2, \dots$ such that $\hat{\Sigma} = \sum_{j\geq 1} \hat{\lambda}_j \hat{u}_j \otimes \hat{u}_j$.

Given a natural number $d \geq 1$, a basic problem is to bound the distance between the eigenspaces $V_d = \operatorname{span}(u_1, \dots, u_d)$ and $\hat{V}_d = \operatorname{span}(\hat{u}_1, \dots, \hat{u}_d)$. Letting P_{V_d} and $P_{\hat{V}_d}$ be the orthogonal projections onto V_d and $\hat{V}_d$, respectively, a natural distance is given by the Hilbert-Schmidt norm $\|P_{\hat{V}_d} - P_{V_d}\|_2$, which has a geometric interpretation in terms of principal angles. The most well-known results in this direction are the Davis-Kahan $\sin\Theta$ theorem and its generalizations, which give upper bounds in terms of the eigenvalue separation and the size of the perturbation, see e.g. [2, 3, 1]. In many cases, more precise bounds can be derived using perturbation theory for linear operators, as developed in [6].

In this work, we establish a local $\sin\Theta$ theorem, tailored for empirical covariance operators. One of the key ingredients is a contraction property, intimately connected to a relative eigenvalue separation condition. A first result in this direction is the following theorem.

Theorem 1. *Let $d \geq 1$. Let $x > 0$ be such that $|\langle u_j, Eu_k\rangle|/\sqrt{\lambda_j \lambda_k} \leq x$ for all $j, k \geq 1$. Suppose that*

$$\sum_{j\leq d} \frac{\lambda_j}{\lambda_j - \lambda_{d+1}} + \sum_{k>d} \frac{\lambda_k}{\lambda_d - \lambda_k} \leq 1/(4x).$$

Then we have

$$\|P_{\hat{V}_d} - P_{V_d}\|_2 \leq 2\sqrt{2}x \sqrt{\sum_{j\leq d} \sum_{k>d} \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2}}.$$

Let us discuss an application of this result to empirical covariance operators. Let X be a random variable taking values in $\mathcal{H}$. Suppose that X is centered and strongly square-integrable, meaning that $\mathbb{E}X = 0$ and $\mathbb{E}\|X\|^2 < \infty$. Let $\Sigma = \mathbb{E}X \otimes X$ be the covariance operator of X, which is a self-adjoint, positive trace class operator. For $j \geq 1$, let $\eta_j = \lambda_j^{-1/2}\langle u_j, X\rangle$ be the j-th Karhunen-Loève coefficient of X. Let $X_1, \dots, X_n$ be n independent copies of X and let

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i \otimes X_i$$

be the empirical covariance operator. Combining Theorem 1 with Burkholder's inequality and Markov's inequality, we get:

Corollary 1. *Let $d \geq 1$. In the above setting, suppose that $\sup_{j\geq 1} \mathbb{E}|\eta_j|^q \leq C_\eta$ for $q \geq 4$ and some constant $C_\eta > 0$. For $t > 0$, suppose that*

$$\frac{t}{\sqrt{n}}\left(\sum_{j\leq d}\frac{\lambda_j}{\lambda_j - \lambda_{d+1}} + \sum_{k>d}\frac{\lambda_k}{\lambda_d - \lambda_k}\right) \leq 1/4.$$

Then, with probability at least $1 - Cp^2 t^{-q/2}$, we have

$$\|P_{\hat{V}_d} - P_{V_d}\|_2^2 \leq \frac{8t^2}{n}\sum_{j\leq d}\sum_{k>d}\frac{\lambda_j\lambda_k}{(\lambda_j - \lambda_k)^2}.$$

Here, p is the dimension of $\mathcal{H}$ and $C > 0$ is a constant which depends only on q and C_η. Moreover, the bound also holds with probability at least $1 - Cd_0^2 t^{-q/2}$, with d_0 such that $\lambda_{d_0} \leq \lambda_d/2$.

References

[1] R. Bhatia, *Matrix analysis*, Springer-Verlag, New York (1997).

[2] C. Davis and W. M. Kahan, *Some new bounds on perturbation of subspaces*, Bull. Amer. Math. Soc. **75** (1969), 863–868.

[3] I. C. F. Ipsen, *An overview of relative* $\sin\Theta$ *theorems for invariant subspaces of complex matrices*, J. Comput. Appl. Math. **123** (2000), 131–153.

[4] M. Jirak and M. Wahl, *Relative perturbation bounds with applications to empirical covariance operators*, available at https://arxiv.org/pdf/1802.02869 (2018).

[5] M. Jirak and M. Wahl, *A tight* $\sin\Theta$ *theorem for empirical covariance operators*, available at https://arxiv.org/pdf/1803.03868 (2018).

[6] T. Kato, *Perturbation theory for linear operators*, Springer-Verlag, Berlin (1995), reprint of the 1980 edition.

Tensor Sparsification and Tensor Completion

Dong Xia

(joint work with Ming Yuan, Cun-Hui Zhang)

We propose a novel tensor sparsification algorithm based on weighted sampling which significantly improves the existed results in sampling complexity. It is shown that our algorithm can be applied for approximating tensor SVD with both space and time complexity.

In addition, we propose two frameworks for estimating a low rank tensor from a subset of its entries with focus on both the statistical and computational efficiencies. In the noiseless setting, we show that a gradient descent algorithm with initial value obtained from a novel spectral method can reconstruct the tensor with sharp sample size requirement. Unlike those earlier approaches for tensor completion, our method is efficient to compute, easy to implement, and does not impose extra structures on the tensor. If the observations are noisy, we show that an even simpler algorithm by combining spectral thresholding and power iterations achieves the optimal rates of convergence which fills in the void of statistical property of noisy tensor completion problems. Even under weak conditions, our algorithm significantly outperforms the existing approaches in the literature.

References

[1] Dong Xia and Ming Yuan, *On Polynomial Time Methods for Exact Low Rank Tensor Completion*, arXiv preprint arXiv:1702.06980, 2017.

[2] Dong Xia and Ming Yuan, *Effective Tensor Sketching via Sparsification*, arXiv preprint arXiv:1710.11298, 2017.

[3] Dong Xia and Ming Yuan and Cun-Hui Zhang, *Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries*, arXiv preprint arXiv:1711.04934, 2017.

Reporter: Anru Zhang