Mathematisches Forschungsinstitut Oberwolfach

Report No. 16/2017

DOI: 10.4171/OWR/2017/16

Statistical Recovery of Discrete, Geometric and Invariant Structures

Organised by
Peter Bühlmann, Zürich
Axel Munk, Göttingen
Martin Wainwright, Berkeley
Bin Yu, Berkeley

19 March – 25 March 2017

Abstract. The main objective of the workshop was to bring together researchers in mathematical statistics and related areas in order to discuss recent advances and problems associated with statistical recovery of geometric and invariant structures. Topics include adaptive estimation, confidence sets and testing techniques, as well as statistical algorithms for geometrical structure recovery and data analysis.

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

Introduction by the Organisers

The theory of optimal function and density estimation has been fundamental to mathematical statistics for many decades. However, it is well understood nowadays that for many modern complex statistical models the full reconstruction of the underlying function is an overly ambitious task. Similarly, this problem arises for the recovery of the full parameter vector in high and ultra high dimensional (generalized) linear models. This becomes particularly apparent if rigorous statistical inference is targeted, e.g. confidence statements for the underlying object to be recovered. There is a variety of results available nowadays which detail these limitations: as one concrete example, it has been shown that adaptive confidence bands (in sup norm) for functions with different smoothness index (e.g. for Sobolev scales) cannot exist as this function space is too complex.

Currently, two possible routes out of this dilemma are being developed, and both have been discussed in this workshop to elucidate a unifying perspective:

(1) To restrict the object of interest by prior information on its geometry, in particular for discrete structures. This leads to new challenges for geometrically-constrained inference at the cutting edge of statistical efficiency and efficient computation. This includes the estimation of discrete structures such as the active nodes in graphical models, networks, or rankings.

(2) To focus on partial information of the structure which potentially allows for inference uniformly over the underlying set of models.

In this workshop a large variety of new results on geometric and invariant inference acquired in seemingly different directions have been reflected from this unifying point of view. The following aspects of geometric and invariant inference jave been in the focus of the proposed meeting.

High dimensional Inference. Various aspects, including optimal shrinkage of covariance matrices (D. Donoho), the relationship between Slope estimation and the Lasso estimator (A. Tsybakov), and compatibility constants for the Lasso (S. van de Geer) have been discussed.
Recovery of Algebraic Structures. P. Rigollet reported on recent results for estimation under algebraic constraints, S. Mukherjee on the geometry of synchronization problems and learning group actions, and M. Yuan on how to analyze large data tensors.
Statistical Optimal Transport. V. Panaretos discussed Fréchet means and procrustes analysis in Wasserstein space and M. Cuturi surveyed recent methods for computation of regularized optimal transport. M. Sommerfeld addressed inference issues for the empirical Wasserstein distance.
Recovery of Networks. S. Olhede surveyed statistical issues of network recovery with combinatorical and nonparametric methods and E. Levina discussed prediction issues in networks with cohesion. In his talk, A. Rinaldo addressed Markov properties of networks models, including attention to their geometry and exchangeability issues.
Inference for Complex Structured Data. E. Candes discussed model-free knockoff methods for replicable selections and S. Balakrishnan discussed local minimax results for hypothesis testing of densities and high-dimensional multinomials. W. Polonik reported on recent results to extract multiscale geometric information extraction and its use for classification.
Causal Inference. T. Richardson discussed how to identify nonparametrically causal effects in the presence of unobserved variables and P. Jonas gave a survey talk on causality and novel ways of exploiting invariance for causal inference.
Multidimensional Regularization. This topic has been addressed by L. Dümbgen, who related geodesic convexity and regularized scatter estimation; by R. Samworth who talked on efficient multivariate entropy estimation; and by V. Vu who prsented recent results on group invariance and computational sufficiency for regularized M-estimators.

Clustering and Classification. V. Spokoiny presented a novel methodology for nonparametric clustering using adaptive weights and H. Zhou talked on statistical and computational guarantees for Lloyd's algorithm and variants thereof. Finally J. Schmidt-Hieber discussed nonparametric Bayesian analysis for support boundary recovery and M. Belkin gave a talk on eigenvectors of orthogonally decomposable functions.

The workshop was complemented by a young researchers late night session, where a total of 11 PhD students and early postdocs presented their work in short talks. This was accompanied by wine and cheese served by the organizers, which created a particularly relaxed atmosphere.

In summary, this was an extremely fruitful and lively workshop where many ideas around geometric and invariant inference have been exchanged between different communities. This includes experts in network analysis, sparse recovery, function estimation, statistics for metric structures and causality, to mention a few.

Acknowledgement: The organizers would like to cordially thank the MFO and its staff for its support and hospitality. 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 their support.

Workshop: Statistical Recovery of Discrete, Geometric and Invariant Structures

Table of Contents

Abstracts

Optimal Shrinkage of Covariance Matrices in light of the spiked covariance model

David Donoho
(joint work with Matan Gavish, Behrooz Ghorbani, and Iain Johnstone)

In recent years, there has been a great deal of excitement about 'big data' and about the new research problems posed by a world of vastly enlarged datasets.

In response, the field of Mathematical Statistics increasingly studies problems where the number of variables measured is comparable to or even larger than the number of observations. Numerous fascinating mathematical phenomena arise in this regime; and in particular theorists discovered that the traditional approach to covariance estimation needs to be completely rethought, by appropriately shrinking the eigenvalues of the empirical covariance matrix.

This talk briefly reviews advances by researchers in random matrix theory who in recent years solved completely the properties of eigenvalues and eigenvectors under the so-called spiked covariance model. By applying these results it is now possible to obtain for the spiked model the exact optimal nonlinear shrinkage of eigenvalues for certain specific measures of performance, as has been shown in the case of Frobenius loss by Nobel and Shabalin, and for many other performance measures by Donoho, Gavish, and Johnstone. Our presentation at Oberwolfach discussed results of [3] on optimal shrinkage for a range of performance 'decomposable' performance measures including operator norm of $\hat{\Sigma} - \Sigma$, Stein Loss, and Frobenius norm of $\hat{\Sigma}^{-1} - \Sigma^{-1}$.

The last part of the talk focused on recent results [4] of the author and Behrooz Ghorbani on optimal shrinkage for the condition number of the relative error matrix; this presents new subtleties as this loss is not decomposable in the sense of [3]. The exact optimal solutions were described, and stylized applications to Muti-User Covariance estimation and Multi-Task Discriminant Analysis were developed.

In more detail, Ghorbani and Donoho considered the following loss function:

$$L(\hat{\Sigma}, \Sigma) = \kappa(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})$$

where $\kappa(\Delta) = \|\Delta\|/\Delta\|^{-1}$ is the condition number. We assumed data $X_i \sim_{iid} N(0, \Sigma)$ $i = 1, \ldots, n$, and $X_i \in \mathbf{R}^p$, and assume $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. We worked in the spiked model asymptotic where $\Sigma = diag(\ell_1, \ldots, \ell_r, 1, 1, \ldots, 1)$ and $\ell_i > 1$, with $r \geq 1$ fixed. Let X denote the n by p data matrix and $\mathbf{S} = \frac{1}{n}X'X$ be the usual empirical second-moment matrix and let $\mathbf{S} = V\Lambda V'$ be its usual eigendecomposition, where V is orthogonal and Λ is diagonal with the ordered eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ along the diagonal. We say that $\hat{\Sigma}$ is *orthogonally equivariant* if $\hat{\Sigma}(U'\mathbf{S}U) = U\hat{\Sigma}(X)U'$ for all $U \in O(p)$. Orthogonally-equivarant procedures are in a certain sense coordinate free.

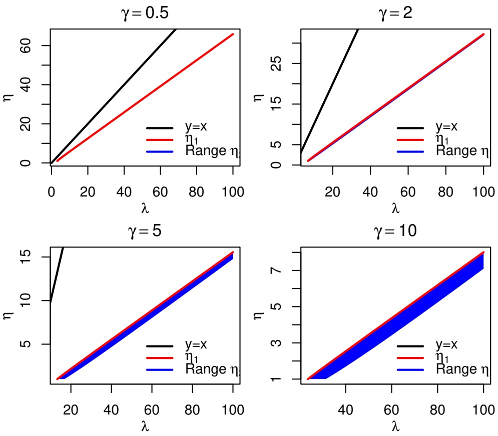

Figure 1.1. Optimal shrinkage in the nonlinear multispike case. There are two parameters: γ and λ_1. The red curve describes the shrinkage function for the top eigenvalue. The blue area is the range of $\eta(\lambda, \lambda_1, \gamma)$ where λ_1 varies from λ to infinity.

Theorem 1. (Optimal Asymptotic Loss) *The following limit exists almost surely:*

$$\lim_{n\to\infty} \inf_{\hat{\Sigma}\in OE} L(\hat{\Sigma}, \Sigma) =_{a.s.} L^*(\ell_1, \dots, \ell_r; \gamma);$$

say. Here the infimum is over orthogonally equivariant procedures. We define in closed form a function κ_1^* depending only on the aspect ratio γ and on the top spike eigenvalue ℓ_1, for which:

$$L^*(\ell_1, \dots, \ell_r; \gamma) = \kappa_1^*(\ell_1; \gamma).$$

Theorem 2. (Asymptotically Optimal Nonlinearity) *We define a closed-form shrinkage nonlinearity $\lambda \mapsto \eta^*(\lambda; \lambda_{1,n}, \gamma)$ having two tuning parameters: $\gamma = p/n$ and $\lambda_{1,n}$ the top empirical eigenvalue. Applying this nonlinearity to each of the empirical eigenvalues (λ_i), produces the diagonal matrix $\eta^*(\Lambda) = diag(\eta^*(\lambda_i))$. These shrunken eigenvalues induce the orthogonally-equivariant covariance estimator $\hat{\Sigma}^* = V\eta^*(\Lambda)V'$; this estimator is asymptotically optimal among orthogonally equivariant procedures under relative condition number loss:*

$$\lim_{n\to\infty} L(\hat{\Sigma}^*, \Sigma) =_{a.s.} \kappa_1^*(\ell_1; \gamma).$$

While the exact closed forms of κ_1^* and η^* are given in our paper, they would take too much space to describe here. The results are quantitatively quite similar to those one would obtain by the generalized soft threshold rule $\eta(\lambda) = 1+b\cdot(\lambda-\lambda_+)_+$, where $\lambda_+ = (1+\sqrt{\gamma})^2$ is the upper edge of the bulk distribution of eigenvalues and $b = 1/(1+\gamma)$. The maximal regret of this thresholding rule as compared to the optimal rule is only a few percent, for $\gamma \le 2$.

References

[1] Baik, Jinho; Ben Arous, Gerard; Peche, Sandrine Phase transition of the largest eigenvalue for non-null complex sample covariance matrices arXiv:math/0403022

[2] Baik, Jinho; Silverstein, Jack W. Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models arXiv:math/0408165

[3] Donoho, David L.; Gavish, Matan; Johnstone, Iain M. Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model arXiv:1311.0851

[4] Donoho, David L.; Ghorbani, Behrooz Optimal Shrinkage of Eigenvalues for relative condition number loss in the Spiked Covariance Model Manuscript.

[5] Shabalin, Andrey; Nobel, Andrew Reconstruction of a Low-rank Matrix in the Presence of Gaussian Noise arXiv:1007.4148

Slope meets Lasso and improved oracle bounds for least squares estimators with convex penalty

Alexandre Tsybakov

We show that two polynomial time methods, a Lasso estimator with adaptively chosen tuning parameter and a Slope estimator, adaptively achieve the exact minimax prediction and ℓ_2 estimation rate $(s/n)\log(p/s)$ in high-dimensional linear regression on the class of s-sparse target vectors in $\mathbf{R}^p$. This is done under the Restricted Eigenvalue (RE) condition for the Lasso and under a slightly more constraining assumption on the design for the Slope. The main results have the form of sharp oracle inequalities accounting for the model misspecification error. The minimax optimal bounds are also obtained for the ℓ_q estimation errors with $1 \leq q \leq 2$ when the model is well-specified. The results are non-asymptotic, and hold both in probability and in expectation. One notable difference from the previous studies of the Lasso and related methods is in the fact that the tuning parameters of the estimators do not depend on the confidence level. In particular, this allows one to derive oracle inequalities in expectation for any moments, which was not possible with the previously known techniques. The assumptions that we impose on the design are satisfied with high probability for a large class of random matrices with independent and possibly anisotropically distributed rows. We give a comparative analysis of conditions, under which oracle bounds for the Lasso and Slope estimators can be obtained. In particular, we show that several known conditions, such as the RE condition and the sparse eigenvalue condition are equivalent if the ℓ_2-norms of regressors are uniformly bounded. Finally, the techniques and the results are extended to the study of more general penalized least squares estimators in a Hilbert space setting, covering as specific cases the group Lasso, and the nuclear norm penalized estimation. The talk is based on a joint work with Pierre C. Bellec and Guillaume Lecué [1]-[3].

References

[1] P.C. Bellec and A.B. Tsybakov. *Bounds on the prediction error of penalized least squares estimators with convex penalty.* `arxiv:1609.06675`

[2] P.C. Bellec, G. Lecué and A.B. Tsybakov. *Slope meets Lasso: Improved oracle bounds and optimality.* `arxiv:1605.08651`

[3] P.C. Bellec, G. Lecué and A.B. Tsybakov. *Towards the study of least squares estimators with convex penalty.* `arxiv:1701.09120`

Some exercises with the Lasso and its compatibility constant

Sara van de Geer

Let $X \in \mathbb{R}^{n\times p}$ be an $n \times p$ matrix and $\beta^0 \in \mathbb{R}^p$ be a fixed vector. We examine the Lasso for the noiseless case

$$\beta^* := \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|X(\beta-\beta^0)\|_2^2 + 2\lambda\|\beta\|_1 \right\}$$

and compare it with the noisy Lasso

$$\hat\beta := \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + 2\lambda\|\beta\|_1 \right\}$$

where $Y = X\beta^0 + \epsilon$ is a noisy version of $X\beta^0$. The prediction error $\|X(\hat\beta - \beta^0)\|_2$ depends on a "bias" term $\|X(\beta^* - \beta^0)\|_2$ and a square-root "$\sqrt{\text{variance}}$" term $\|X(\hat\beta - \beta^*)\|_2$. Theorem 1 below shows that under some conditions the "bias" term is the dominating one. The Gram matrix is $\hat\Sigma := X^T X$. Let Σ_0 be some (other) $p \times p$ matrix Σ_0 (for instance in the case of random design Σ_0 could be the expected value $\mathbb{E}\hat\Sigma$ of $\hat\Sigma$). The largest eigenvalue of Σ_0 is denoted by $\Lambda_{\max}(\Sigma_0)$.

Theorem 1. *Let $0 < \alpha < 1$ and $0 < \alpha_1 < 1$ be fixed, write $\lambda_0 := \sqrt{2\log(2p/\alpha)/n}$ and let $\eta\lambda > \lambda_0$ for some $0 \le \eta < 1$.*
Suppose
- the columns of X are normalized to have length at most 1,
- the noise ϵ consists of i.i.d. Gaussians with mean zero and variance $1/n$,
- the matrices $\hat\Sigma$ and Σ_0 are close enough in the sense that

$$\xi := \|\hat\Sigma - \Sigma_0\|_\infty \|\beta^* - \beta^0\|_1 < \lambda(1-\eta).$$

Then with probability at least $1 - \alpha - \alpha_1$

$$\|X(\hat\beta - \beta^*)\|_2 \le \frac{\Lambda_{\max}^{1/2}(\Sigma_0)\biggl(\|X(\beta^*-\beta^0)\|_2^2 + \xi\|\beta^* - \beta^0\|_1 \biggr)^{1/2}}{\biggl(\lambda(1-\eta) - \xi\biggr)} + \sqrt{\frac{2\log(1/\alpha_1)}{n}}.$$

Next, we study the noiseless Lasso. For $S \subset \{1, \ldots, p\}$, and $\beta \in \mathbb{R}^p$, the vector β_S denotes the vector β with its entries in the set S set to zero. Moreover, we write $\beta_{-S} := \beta - \beta_S$. The compatibility constant is

$$\hat{\phi}^2(S) := \min\left\{ |S| \|X\beta\|_2^2 : \ \|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \leq 1 \right\}.$$

Then

$$\|X(\beta^* - \beta^0)\|_2^2 + 2\lambda \|\beta^*_{-S_0}\|_1 \leq \frac{\lambda^2 |S_0|}{\hat{\phi}^2(S_0)},$$

where S_0 is the support set of β^0 (see for example [1] and its references). To see whether this bound is tight, one may study several particular cases.

Lemma 1. *Suppose that*

$$\hat{\Sigma} = \begin{pmatrix} 1 & -\hat{\rho}_1 & & & \\ -\hat{\rho}_1 & 1 & & & \\ & & \ddots & & \\ & & & 1 & -\hat{\rho}_N \\ & & & -\hat{\rho}_N & 1 \end{pmatrix}.$$

where $0 \leq \hat{\rho}_k < 1$ *for all* $k = 1, \ldots, N$. *Let* $S_0 = \{1, \ldots, p\}$ $(p = 2N)$. *Write* $\hat{\varphi}_k^2 := 1 - \hat{\rho}_k$ *and assume that* $\beta^0_{2k-1} \geq \beta^0_{2k} \geq \lambda / \hat{\varphi}_k^2$ *for each* k. *Then it holds that*

$$\|X(\beta^* - \beta^0)\|_2^2 = \frac{\lambda^2 |S_0|}{\hat{\phi}^2(S_0)},$$

and

$$\hat{\phi}^2(S_0) = \frac{N}{\sum_{k=1}^N 1/\hat{\varphi}_k^2}.$$

Thus in the situation of Lemma 1 the upper bound and lower bound match. Note that the compatibility constant can be much larger than the minimal eigenvalue of $\hat{\Sigma}$ (which is $\min_k \hat{\varphi}_k^2$ in the case of Lemma 1).

In Lemma 1 there are no inactive variables. One can consider various scenario's with inactive variables. One such (simple) scenario is the following.

Lemma 2. *Suppose that*

$$\hat{\Sigma} = \begin{pmatrix} 1 & -\hat{\rho} & C\hat{\varphi}^2/2 \\ -\hat{\rho} & C\hat{\varphi}^2/2 & \\ C\hat{\varphi}^2/2 & C\hat{\varphi}^2/2 & 1 \end{pmatrix}$$

where $0 \leq \hat{\rho} < 1$, $\hat{\varphi}^2 := 1 - \hat{\rho}$ *and* $1 < C < 2/\hat{\varphi}^2$. *Let* $\hat{\tau}^2 := 1 - C^2\hat{\varphi}^2/2$. *Assume that* $\beta^0_1 \geq \beta^0_2 \geq \lambda C(C-1)/(2\hat{\tau}^2)$ *and* $\beta^0_3 = 0$. *Then*

$$\|X(\beta^* - \beta^0)\|_2^2 + 2\lambda \|\beta^*_{-S_0}\|_1 = \frac{\lambda^2 |S_0|}{\hat{\phi}^2(S_0)} - \frac{\lambda^2}{\hat{\tau}^2}.$$

References

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

Optimal rates of estimation for the multi-reference alignment problem

Philippe Rigollet
(joint work with Afonso S. Bandeira, Jonathan Weed)

The multi-reference alignment problem and its variants arise in various scientific and engineering applications such as structural biology [SVN$^+$05, TS12, Sad89], image recognition [Bro92], and signal processing [ZvdHGG03]. A striking feature of this class of problems is that each observation is not only observed in a noisy setting but is also altered by an latent transformation that reflects underlying heterogeneity of the data. The precise nature of this transformation depends on the specific application, but it can often be characterized as the action of the unknown element of a known group.

Concretely, one observes n independent vectors $Y_1, \ldots, Y_n \in \mathbb{R}^d$ given by

$$Y_i = R_i\theta + \sigma\xi_i\,,$$

where θ is an unknown parameter of interest, $\xi_i \sim \mathcal{N}(0, I_d)$ is independent Gaussian noise, and R_i is an unknown element from a known compact subgroup $\mathcal{G}$ of the orthogonal group in d dimensions. The multi-reference alignment problem [BCSZ14] takes $\mathcal{G}$ to be the group of cyclic shifts of the coordinates of θ: given $\ell \in [d]$, the jth coordinate of the vector $R_\ell\theta$ is given by $(R_\ell\theta)_j = \theta_{j+\ell \pmod d}$. This group has a simple representation in the Fourier domain, where it acts on the phases of the Fourier coefficients.

We focus on a slightly different group isomorphic to the *circle group* $U(1)$ of unit-norm complex numbers, which is both slightly easier to analyze and better corresponds to the situation in practice. By analogy with the action of the group of cyclic shifts on the phases of the Fourier coefficients, it is most convenient to define this group in the Fourier domain. Given $z \in U(1)$ we define the operator R_z on $\mathbb{R}^d$ by its action on the Fourier transform $\hat{\theta}$:

$$\widehat{R_z\theta}_j = z^j\hat{\theta}_j \quad \text{for } -\lfloor d/2 \rfloor \leq j \leq \lfloor d/2 \rfloor.$$

We define the group of such operators by $\mathcal{F}$ and call R_z a *fractional cyclic shift.* For the sake of exposition, in the sequel, we will focus on $\mathcal{F}$ and omit the adjective "fractional" when referring to shifts.

We propose to analyze the multi-reference alignment problem as a continuous mixture of Gaussians. Since the Gaussian distribution is invariant under cyclic shift, we can without loss of generality assume that the shifts R_i are independent and uniformly (i.e., according to the Haar probability measure) distributed over $\mathcal{F}$.

We are interested in recovering the unknown parameter θ, which we can clearly only do up to cyclic shift. Define the pseudo-metric ρ on $\mathbb{R}^d$ by

$$\rho(\theta, \phi) = \min_{R \in \mathcal{F}} \|\theta - R\phi\|\,.$$

We assume throughout that the noise variance σ^2 is known, and we are interested in signals θ for which a parametric rate is achievable. Hence we assume the existence of an absolute constant c not depending on n such that $c^{-1} \leq |\hat{\theta}_j| \leq c$ for all j such that $\hat{\theta}_j \neq 0$, and denote the set of such vectors by $\mathcal{T}$. Surprisingly, even under this assumption, the multi-reference alignment problem suffers from the curse of dimensionality.

Theorem 1. *Let $2 \leq s \leq \lfloor d/2 \rfloor$. Let $\mathcal{T}_s$ be the set of vectors $\theta \in \mathcal{T}$ such that the support of $\hat{\theta}$ lies in $\{-s, \dots, s\}$. Then,*

$$\inf_{T_n} \sup_{\theta \in \mathcal{T}_s} \mathbb{E}_\theta[\rho(T_n, \theta)] \asymp \frac{\sigma^{2s-1}}{\sqrt{n}}(1 + o_n(1)),$$

where the infimum is taken over all estimators T_n of θ and where the symbol $\asymp$ hides constants depending on d but on no other parameter. A modified MLE achieves this rate.

The worst case bounds appearing in Theorem 1 contrast sharply with the typical case: in a companion paper [PWB$^+$17], we show that signals θ whose Fourier transform has full support may be estimated as the same rate as signals in $\mathcal{T}_2$.

The proof of Theorem 1 relies on tight control of the Kullback-Leibler divergence between two probability distributions via their moment tensors. Our main technical result makes this connection precise. Note that this Theorem holds for any compact subgroup of the orthogonal group and is not specific to cyclic shits.

Theorem 2. *Let θ be a fixed vector in $\mathbb{R}^d$ such that $\|\theta\| \leq 1$. Let $\phi \in \mathbb{R}^d$ be such that $\rho(\theta, \phi) = \varepsilon \leq \|\theta\|$. Let R be a random element drawn according to the Haar probability measure on any compact subgroup $\mathcal{G}$ of the orthogonal group in d dimensions. For all $m \geq 1$, let $\Delta_m = \mathbb{E}[(R\theta)^{\otimes m} - (R\phi)^{\otimes m}]$. If there exists $k \geq 1$ such that, as $\varepsilon \to 0$,*

$$\|\Delta_m\| = o(\varepsilon) \text{ for } m = 1, \dots, k-1, \quad \text{and} \quad \|\Delta_k\| = \Omega(\varepsilon),$$

then $\|\Delta_k\| = \Theta(\varepsilon)$. Moreover, for $\sigma \geq 1$ there exist universal constants c and $\bar{C}$ and constant $\underline{C}_d$ that depends only on d, all positive and such that

$$\frac{c^k}{\sigma^{2k} k!}\|\Delta_k\|^2 - \underline{C}_d \frac{\varepsilon^2}{\sigma^{2k+2}} \leq D(\theta \parallel \phi) \leq \frac{2}{\sigma^{2k} k!}\|\Delta_k\|^2 + \bar{C}\frac{\varepsilon^2}{\sigma^{2k+2}}.$$

In particular, there exists positive σ_0, ε_0 that depend on d such that for all $\sigma \geq \sigma_0$, and θ, ϕ such that $\|\theta\| \leq 1, \rho(\theta, \phi) \leq \varepsilon_0$, it holds

$$D(\theta \parallel \phi) \asymp \sigma^{-2k} \rho^2(\theta, \phi),$$

where the symbol $\asymp$ hides constants depending on d but on no other parameters.

Theorem 2 immediately implies minimax lower bounds via LeCam's method [LeC73]. On the other hand, we also show how to transform lower bounds on $D(\mathrm{P}_\theta \parallel \mathrm{P}_\phi)$ into *uniform* upper bounds on the performance of the MLE. Note that this analysis departs from the classical *pointwise* rate of convergence for MLE that guarantees a rate of convergence $n^{-1/2}$ *for each fixed choice of parameter* as $n \to \infty$. Our tools strengthen this result considerably. Indeed, we show that for reasonable choices of θ, the MLE achieves a rate of $n^{1/2}$ *uniformly* over all choices of θ, with an optimal dependence on the noise level σ.

To prove the lower and upper bounds in Theorem 1 we use Fourier-theoretic arguments to show that, if $\theta \in \mathcal{T}_s$, then for any ϕ with the same support as θ such that $\rho(\theta, \phi) = \varepsilon$, there exists $k \leq 2s - 1$ such that $\|\Delta_k\| = \Omega(\varepsilon)$. Conversely, we show how to exhibit vectors θ and ϕ in $\mathcal{T}_s$ for which $\|\Delta_m\| = 0$ for all $m < 2s - 1$. These two results, combined with Theorem 2, imply tight lower and upper bounds on on $D(\mathrm{P}_\theta \parallel \mathrm{P}_\phi)$ over the class $\mathcal{T}_s$. LeCam's method and our new analysis of the maximum likelihood estimator then imply Theorem 1.

References

[BCSZ14] Afonso S. Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science*, pages 459–470. ACM, New York, 2014.

[Bro92] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.

[LeC73] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[PWB+17] A. Perry, J. Weed, A. S. Bandeira, P. Rigollet, and A. Singer. The sample complexity of multi-reference alignment. Manuscript, 2017.

[Sad89] B. M. Sadler. Shift and rotation invariant object reconstruction using the bispectrum. In *Workshop on Higher-Order Spectral Analysis*, pages 106–111, Jun 1989.

[SVN+05] Sjors H.W. Scheres, Mikel Valle, Rafael Nuñez, Carlos O.S. Sorzano, Roberto Marabini, Gabor T. Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of Molecular Biology*, 348(1):139 – 149, 2005.

[TS12] D. L. Theobald and P. A. Steindel. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics*, 28(15):1972–1979, 2012.

[ZvdHGG03] J. P. Zwart, R. van der Heiden, S. Gelsema, and F. Groen. Fast translation invariant classification of HRR range profiles in a zero phase representation. *Radar, Sonar and Navigation, IEE Proceedings*, 150(6):411–418, 2003.

The Geometry of Synchronization Problems and Learning Group Actions

Sayan Mukherjee

(joint work with Tingran Gao, Jacek Brodzki)

We develop a geometric framework that characterizes the synchronization problem — the problem of consistently registering or aligning a collection of objects. The theory we formulate characterizes the cohomological nature of synchronization based on the classical theory of fibre bundles. We first establish the correspondence between synchronization problems in a topological group Gover a connected graph Γ and the moduli space of flat principal G-bundles over Γ, and develop a discrete analogy of the renowned theorem of classifying flat principal bundles with fix base and structural group using the representation variety. In particular, we show that prescribing an edge potential on a graph is equivalent to specifying an equivalence class of flat principal bundles, of which the triviality of holonomy dictates the synchronizability of the edge potential. We then develop a twisted cohomology theory for associated vector bundles of the flat principal bundle arising from an edge potential, which is a discrete version of the twisted cohomology in differential geometry. This theory realizes the obstruction to synchronizability as a cohomology group of the twisted de Rham cochain complex. We then build a discrete twisted Hodge theory — a fibre bundle analog of the discrete Hodge theory on graphs — which geometrically realizes the graph connection Laplacian as a Hodge Laplacian of degree zero. Motivated by our geometric framework, we study the problem of learning group actions — partitioning a collection of objects based on the local synchronizability of pairwise correspondence relations. A dual interpretation is to learn finitely generated subgroups of an ambient transformation group from noisy observed group elements. A synchronization-based algorithm is also provided, and we demonstrate its efficacy using simulations and real data.

References

[1] T. Gao, J. Brodzki, S. Mukherjee, *The Geometry of Synchronization Problems and Learning Group Actions*, arXiv:1610.09051 (2017).

On Polynomial Time Methods for Low Rank Tensor Completion

Ming Yuan
(joint work with Dong Xia)

Let $\mathbf{T} \in \mathbb{R}^{\mathbf{d_1}\times\cdots\times\mathbf{d_k}}$ be a kth order tensor. The goal of tensor completion is to recover $\mathbf{T}$ based on a subset of its entries $\{T(\omega) : \omega \in \Omega\}$ for some $\Omega \subset [d_1] \times \cdots \times [d_k]$ where $[d] = \{1, 2, \ldots, d\}$. The problem of tensor completion has attracted a lot of attention in recent years due to its wide range of applications. In particular, the second order (matrix) case has been extensively studied. One of the main revelations from these studies is that, although the matrix completion problem is in general NP-hard, it is possible to develop tractable algorithms to achieve exact recovery with high probability. Naturally one asks if the same can be said for higher order tensors. This seemingly innocent task of generalizing from second order to higher order tensors turns out to be rather delicate.

The challenges in dealing with higher order tensors comes from both computational and theoretical fronts. On the one hand, many of the standard operations for matrices become prohibitively expensive to compute for higher order tensors. A notable example is the computation of tensor spectral norm. For second order tensors, or matrices, the spectral norm is merely its largest singular value and can be computed with little effort. Yet this is no longer the case for higher order tensors where computing the spectral norm is NP-hard in general. On the other hand, many of the mathematical tools, either algebraic such as characterizing the subdifferential of the nuclear norm or probabilistic such as concentration inequalities, essential to the analysis of matrix completion are still under development for higher order tenors. There is a fast growing literature to address both issues and much progresses have been made in both fronts in the past several years.

When it comes to higher order tensor completion, an especially appealing idea is to first unfold a tensor to a matrix and then treat it using techniques for matrix completion. As shown recently, these approaches, although easy to implement, may require an unnecessarily large amount of entries to be observed to ensure exact recovery. As an alternative, nuclear norm minimization can recover a $d \times d \times d$ tensor with multilinear ranks (r, r, r) with high probability with as few as $O((r^{1/2}d^{3/2} + r^2 d)(\log d)^2)$ observed entries. Perhaps more surprisingly, it was later showed that the dependence on d (e.g., the factor $d^{3/2}$) remains the same for higher order tensors and we can reconstruct a kth order cubic tensor with as few as $O((r^{(k-1)/2}d^{3/2} + r^{k-1}d)(\log d)^2)$ entries for any $k \geq 3$ when minimizing a more specialized nuclear norm devised to take into account the incoherence. These sample size requirement drastically improve those based on unfolding which typically require a sample size of the order $r^{\lfloor k/2 \rfloor}d^{\lceil k/2 \rceil}\text{polylog}(d)$. Although both nuclear norm minimization approaches are based on convex optimization, they are also NP hard to compute in general. Many approximate algorithms have also been proposed in recent years with little theoretical justification. It remains unknown if there exist polynomial time algorithms that can recover a low rank tensor exactly with similar sample size requirements. The goal of the present article is to fill in the gap between these two strands of research by developing a computationally efficient approach with tight sample size requirement for completing a third order tensor.

In particular, we show that there are polynomial time algorithms that can reconstruct a $d_1 \times d_2 \times d_3$ tensor with multilinear ranks (r_1, r_2, r_3) from as few as

$$O\Big(r_1r_2r_3(rd_1d_2d_3)^{1/2}\log^{7/2} d + (r_1r_2r_3)^2 rd\log^6 d\Big)$$

entries where $r = \max\{r_1, r_2, r_3\}$ and $d = \max\{d_1, d_2, d_3\}$. This sample size requirement matches those for tensor nuclear norm minimization in terms of its dependence on the dimension d_1, d_2 and d_3 although it is inferior in terms of its dependence on the ranks r_1, r_2 and r_3. This makes our approach especially attractive in practice because we are primarily interested in high dimension (large d) and low rank (small r) instances. In particular, when $r = O(1)$, our algorithms can recover a tensor exactly based on $O(d^{3/2}\log^{7/2} d)$ observed entries, which is nearly identical to that based on nuclear norm minimization.

It is known that the problem of tensor completion can be cast as optimization over a direct product of Grassmannians. The high level idea behind our development is similar to those used earlier for matrix completion: if we can start with an initial value sufficiently close to the truth, then a small number of observed entries can ensure the convergence of typical optimization algorithms on Grassmannians such as gradient descent to the truth. Yet the implementation of this strategy is much more delicate and poses significant new challenges when moving from matrices to tensors.

At the core of our method is the initialization of the linear subspaces in which the fibers of a tensor reside. In the matrix case, a natural way to do so is by singular value decomposition, a tool that is no longer available for higher order tensors. An obvious solution is to unfold tensors into matrices and then applying the usual singular value decomposition based approach. This, however, requires an unnecessarily large sample size. To overcome this problem, we propose an alternative approach to estimating the singular spaces of the matrix unfoldings of a tensor. Our method is based on a carefully constructed estimate of the second moment of appropriate unfolding of a tensor, which can be viewed as a matrix version U-statistics. We show that the eigenspace of the proposed estimate concentrates around the true singular spaces of the matrix unfolding more sharply than the usual singular value decomposition based approaches, and therefore leads to consistent estimate with tighter sample size requirement.

The fact that there exist polynomial time algorithms to estimate a tensor consistently, not exactly, with $O(d^{3/2}\text{polylog}(r, \log d))$ observed entries was first recognized recently based on sum-of-square relaxations of tensor nuclear norm. Although polynomial time solvable in principle, their method requires solving a semidefinite program of size $d^3 \times d^3$ and is not amenable to practical implementation. In contrast, our approach is essentially based on the spectral decomposition of a $d \times d$ matrix and can be computed fairly efficiently.

Once a good initial value is obtained, we consider reconstructing a tensor by optimizing on a direct product of Grassmannians locally. To this end, we consider a simple gradient descent algorithm adapted for our purposes. The main architect of our argument is similar to those for matrix completion. We argue that the objective function, in a suitable neighbor around the truth and including the initial value, behaves like a parabola. As a result, the gradient descent algorithm necessarily converges locally to a stationary point. We then show that the true tensor is indeed the only stationary point in the neighborhood and therefore the algorithm recovers the truth. To prove these statements for higher order tensors however require a number of new probabilistic tools for tensors, and we do so by establishing several new concentration bounds.

Fréchet Means and Procrustes Analysis in Wasserstein Space

Victor M. Panaretos
(joint work with Yoav Zemel)

We consider three interlinked problems at the intersection of functional and geometrical data analysis, involving distributions of finite variance:

(1) Constructing the optimal multicoupling of marginal distributions;

Let $\Lambda_1, \ldots, \Lambda_n$ be measures on $\mathbb{R}^d$. Construct random variables $\{X_i\}_{i=1}^n$ such that $\{X_i \sim \Lambda_i; i = 1, ..., n\}$ and $\sum_{i=1}^n \sum_{j=i+1}^n \mathbb{E}|X_i - X_j|^2 \leq \sum_{i=1}^n \sum_{j=i+1}^n \mathbb{E}|Y_i - Y_j|^2$ for any other collection of random variables $Y_i \sim \Lambda_i$, $i = 1, ..., n$.

(2) Determining the Fréchet mean of random probability measures;

Let $\Lambda_1, \ldots, \Lambda_n$ be measures on $\mathbb{R}^d$. Find the minimiser of the functional $\gamma \mapsto \frac{1}{n}\sum_{i=1}^n d_W^2(\gamma, \Lambda_i)$ over probability measures γ on $\mathbb{R}^d$, *where* $d_W(\Lambda_1, \Lambda_2) = \inf_{X_1 \sim \Lambda_1, X_2 \sim \Lambda_2} \sqrt{\mathbb{E}\|X_1 - X_2\|^2}$ *is the quadratic Wasserstein distance, a.k.a. the Fréchet distance.*

(3) Registering collections of randomly deformed point processes.
Let $\Pi_1, \dots, \Pi_n$ *be* iid *point processes on* $[0,1]^d$*, with mean measure* $\mu = \tau\lambda$*, for* λ *a probability measure and* $\tau > 0$*. Letting* $T_i : [0,1]^d \to [0,1]^d$ *be unobservable random homeomorphisms, recover* $\{\Pi_i\}_{i=1}^n$ *and* $\{T_i\}_{i=1}^n$ *from observation of the* n *warped processes* $\widetilde{\Pi}_i = (T_i)\#\Pi_i$*.*

The common thread is that all three problems are canonically related to the optimal transportation geometry of the Wassertein space of measures on $\mathbb{R}^d$. And, consequently, the solution of these problems and corresponding nonparametric statistical properties bifurcate according to whether $d = 1$ or $d > 1$.

The case $d = 1$. In the one-dimensional case, Wasserstein space is flat, and the solutions of (1) and (2) are straightforward by means of the probability integral transform, at least for absolutely continuous measures. For (1), observe that each term $\mathbb{E}\|X_i - X_j|^2$ can be minimised by defining $X_i = F_{\Lambda_i}^{-1}(U) \sim \Lambda_i$, for a common uniform random variable U on $[0,1]$. Given this, (2) reduces to finding the minimiser of $\sum_{i=1}^n d_W^2(\gamma, \Lambda_i) = \sum_{i=1}^n \mathbb{E}|F_{\Lambda_i}^{-1}(U) - F_\gamma^{-1}(U)|^2 = \sum_{i=1}^n \|F_{\Lambda_i}^{-1} - F_\gamma^{-1}\|_{L^2}^2$, which is attained uniquely at the measure $\bar{\Lambda}$ with quantile function $F_{\bar{\Lambda}}^{-1} = n^{-1}\sum_{i=1}^n F_{\Lambda_i}^{-1}$.

To solve (3), we use the structure of (1) and (2). We show that under the canonical assumptions that $\mathbb{E}[T_i(x)] = x$ (unbiased deformations) and that T is increasing (no time pause or reversal), the structural mean measure λ can be uniquely identified as the *population* Fréchet mean of the random measure $\Lambda = T\#\lambda$, i.e. the unique minimiser of $\gamma \mapsto \mathbb{E}d_W^2(\gamma, T\#\lambda)$; the unobservable T_i are then uniquely identified as the optimal maps from λ to Λ_i.

We then show how λ, $\{\Pi_i\}_{i=1}^n$ and $\{T_i\}_{i=1}^n$ can be recovered non-parametrically by exploiting this structure: first one estimates each $\Lambda_i = (T_i)\#\lambda$ separately by smoothing $\tilde{\Pi}_i$, obtaining $\hat{\Lambda}_i$ (say); then, one estimates λ by the Fréchet–Wasserstein mean of the $\hat{\Lambda}_i$, say $\hat{\lambda}$; and, finally, one estimates T_i as the optimal coupling $\hat{T}_i$ of $\hat{\lambda}$ to each $\hat{\Lambda}_i$, registering the point processes by taking $(\hat{T}_i^{-1})\#\widetilde{\Pi}_i$. The resulting estimators can be shown to be consistent as τ and n diverge. In the case where the $\{\Pi_i\}$ are Poisson processes, convergence rates are obtained and shown to be essentially optimal, and a tangent space central limit theorem is also obtained.
For more details, see Panaretos & Zemel [2].

The case $d > 1$. In this case, Wasserstein space is positively curved, and neither (1) nor (2) can be solved explicitly. For (2), Agueh & Carlier [1] show that a unique Fréchet mean *will* exist, when assuming that the $\{\Lambda_i\}$ are diffuse and with bounded density. They also show that if (1) can be solved, yielding optimally coupled random variables $\{X_1, ..., X_n\}$, then the law of $n^{-1}\sum_{i=1}^n X_i$ will yield the Fréchet mean of $\{\Lambda_1, ..., \Lambda_n\}$. We go in the opposite direction. Assuming that the measures are diffuse and of bounded density, we show that if $\bar{\Lambda}$ is the

Fréchet mean of $\{\Lambda_1, ..., \Lambda_n\}$, and if $Z \sim \bar{\Lambda}$, then $X_i := \mathbf{t}_{\bar{\Lambda}}^{\Lambda_i}(Z)$ yields the optimal multicoupling solution to (1), where $\mathbf{t}_\mu^\nu$ denotes the optimal coupling map between diffuse measures μ and ν, i.e.

$$\mathbf{t}_\mu^\nu \# \mu = \nu \quad \& \quad d_W^2(\mu, \nu) = \int_{\mathbb{R}^d} \|\mathbf{t}_\mu^\nu(x) - x\|^2 \mu(dx).$$

Thus (1) is solved, as long as (2) can be solved. To solve (2), we reduce the problem of finding the Fréchet mean to the solution of successive pairwise optimal transportation problems, a Procrustes analysis heuristic. Using the tangent bundle structure of Wasserstein space, we determine the Fréchet derivative of the functional $F(\gamma) = n^{-1} \sum_{i=1}^n d_W^2(\gamma, \Lambda_i)$, and determine the optimal step-size for a steepest descent minimisation. It turns out that $F'(\gamma) = -n^{-1} \sum_{i=1}^n \left(\mathbf{t}_\gamma^{\Lambda_i} - \mathbf{i}\right)$, for $\mathbf{i}$ the identity map, $\mathbf{i}(x) = x$, and that the optimal step-size is 1. This gives rise to the steepest descent algorithm:

(A) Set a tolerance threshold $\epsilon > 0$.

(B) For $j = 0$, let γ_j be an arbitrary diffuse measure.

(C) For $i = 1, \ldots, n$ solve the (pairwise) Monge problem and find the optimal transport map $\mathbf{t}_{\gamma_j}^{\Lambda_i}$ from γ_j to Λ_i.

(D) Define the map $T_j = n^{-1} \sum_{i=1}^n \mathbf{t}_{\gamma_j}^{\Lambda_i}$.

(E) Set $\gamma_{j+1} = T_j \# \gamma_j$, i.e. push-forward γ_j via T_j to obtain γ_{j+1}.

(F) If $\|F'(\gamma_{j+1})\| < \epsilon$, stop, and output γ_{j+1} as the approximation of $\bar{\Lambda}$ and $\mathbf{t}_{\gamma_{j+1}}^{\Lambda_i}$ as the approximation of $\mathbf{t}_{\bar{\Lambda}}^{\Lambda_i}$, $i = 1, \ldots, n$. Otherwise, go to (C).

We show Wasserstein convergence of the algorithm iterates γ_j to a critical point of F, and provide sufficent conditions for this critical point to be the unique Fréchet mean $\bar{\Lambda}$. In this case, we also deduce uniform convergence of the optimal maps $\mathbf{t}_{\gamma_j}^{\Lambda_i}$ to the optimal multicoupling maps $\mathbf{t}_{\bar{\Lambda}}^{\Lambda_i}$. The algorithm can be seen to be a Procrustes algorithm: at each step it optimally couples every measure to the current iterate, pairwise. It then averages the registration maps, and pushes forward the current iterate by the average registration map. The value of this Procrustean structure goes beyond aesthetics: contrary to the optimal multicoupling problem, pairwise optimal coupling problems can be solved numerically, and thus both the determination of a Fréchet mean, as well as the optimal multicoupling problem are reduced to the computationally feasible pairwise problem of optimal transportation, just as in the one-dimensional case (albeit without closed-form expressions).

With the solution of (1) and (2) under our belt, we then attack problem (3). In the case $d > 1$, we show that under the analogous canonical assumptions on the deformations T_i, the structural mean measure λ can again be uniquely identified as the *population* Fréchet mean of the random measure $\Lambda = T\#\lambda$. The main difference here is that T_i need to be gradients of convex functions rather than increasing maps (Brenier's characterisation). The resulting structure of the problem is analogous to the 1-dimensional case, and estimators can be defined in a similar way – using multivariate optimal transportation and Procrustes analysis for their solution. Though rates of convergence remain elusive, we still deduce consistency of all the nonparametric estimators involved. While the results parallel the one-dimensional case, their derivation requires entirely different techniques, related to the structure of gradients of convex maps and their weak convergence in $\mathbb{R}^d$.

For more details see Zemel & Panaretos [4, 5] and for a detailed presentation, including an introduction to optimal transportation, we refer to the forthcoming monograph by Panaretos & Zemel [3] and the PhD thesis of Zemel [6]. We gratefully acknowledge support by an ERC Starting Grant Award to Victor M. Panaretos.

References

[1] M. Agueh & G. Carlier, *Barycenters in the Wasserstein space*, Society for Industrial and Applied Mathematics, 43 (2): 904–924, 2011.

[2] V.M. Panaretos & Y. Zemel *Amplitude and Phase Variation of Point Processes*, Annals of Statistics 44(2): 771–812, 2016.

[3] V.M. Panaretos & Y. Zemel *Foundations of Statistics in the Wasserstein Space*, in preparation.

[4] Y. Zemel & V.M. Panaretos, *Fréchet Means in Wasserstein Space: Gradient Descent and Procrustes Analysis*, Technical Report #1-16, Chair of Mathematical Statistics, EPFL, http://smat.epfl.ch/reports/1-16.pdf, February 2016.

[5] Y. Zemel & V.M. Panaretos, *Fréchet Means and Procrustes Analysis in Wasserstein Space*, arXiv:1701.06876.

[6] Y. Zemel *Fréchet Means in Wasserstein Space: Theory and Algorithms*, PhD Thesis, École Polytechnique Fédérale de Lausanne, February 2017.

Review of Regularized Optimal Transport

Marco Cuturi

Monge [12] and later Kantorovich [11] built using relatively simple mathematical blocks what is known today as optimal transport theory, a field that has drawn in recent decades the interest of pure and applied mathematicians [10, 16, 14]. One of the most notable results of that theory lies in the definition of a versatile distance between probability measures, known as the Wasserstein distance. Because probability measures are widely used to model social and natural phenomena, the toolbox of optimal transport has been increasingly adopted in a wide array of applied fields, such as economics [9], fluid dynamics [4], quantum chemistry [5, 8], computer vision [13], or graphics [15]. Closer to the audience of this workshop, ideas related to optimal transport are also increasingly adopted by statisticians. The main motivation behind the work presented in this seminar was that I wanted to study new machine learning methodologies built upon the optimal transport geometry. The main obstacle to this goal was computational, since optimal transport is notoriously costly to compute. To avoid that issue, I have proposed 3 years ago [6] a very efficient numerical scheme to solve optimal transport problems that can scale up to large scales and use recent progresses in hardware, namely GPGPUs. This breakthrough has inspired several works in the span of 3 years, in machine learning [17] and beyond [2, 7, 3], which I tried to survey in this talk, with a special emphasis on the problems of Wasserstein barycenters [1] and minimum Wasserstein distance estimation.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative 'Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[3] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *journal=ACM Transactions on Graphics, volume=35, number=4, year=2016*.

[4] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[5] Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.

[6] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[7] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*.

[8] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[9] A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[10] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[11] L. V. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk. USSR*, 37:199–201, 1942.

[12] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[13] Y. Rubner, L.J. Guibas, and C. Tomasi. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proc. of ARPA Image Understanding Works.*, pages 661–668, 1997.

[14] F. Santambrogio. *Optimal Transport for Applied Mathematicians - Calculus of Variations, PDEs and Modeling*. Springer, 2016.

[15] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015)*, 2015.

[16] C. Villani. *Optimal transport: old and new*, volume 338. Springer Verlag, 2009.

[17] C. Zhang, C. Frogner, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems 29*, 2015.

Distributional Limits for Wasserstein Distance on Discrete Spaces

Max Sommerfeld

(joint work with Axel Munk, Carla Tameling, Jörn Schrieber)

The empirical Wasserstein distance between distributions is an attractive tool for statistical applications but suffers from two major obstacles: First, inference is hindered by the lack of distributional limits for spaces other than the real line. Second, the computational cost is prohibitive even for moderately sized problems. We argue that both obstacles can be overcome in the setting of finite spaces. To this end, for probability measures supported on finitely many points, we derive the asymptotic distribution of the Wasserstein distance of empirical distributions as the optimal value of a linear program with random objective function. As a consequence statistical inference for sample based Wasserstein distances becomes doable in large generality. We introduce the concept of directional Hadamard differentiability in this context. To approximate the limiting distribution, we discuss bootstrapping schemes accounting for the non-linear derivative of the Wasserstein distance and explore modifications that reduce the computational burden.

Nevertheless, when problem sizes become large, exact computation of the Wasserstein distance as well as the bootstrap become computationally infeasible. To facilitate inference (e.g. testing) in these situations, we lower bound the Wasserstein distance and stochastically upper bound the limiting distribution using tree metrics and give efficient algorithms to compute these bounds.

For non-inferential tasks such as classification, which often require the fast computation of a large number of Wasserstein distances but may permit an error in each calculation, we propose a probabilistic approximation with exact solvers. The problem size is reduces dramatically by considering only a sub-sample of the original measures and computing the exact distance between these. This scheme is can easily be tuned towards higher accuracy or lower computational burden, works with any solver as a back-end, including entropy regularized versions and comes with non-asymptotic guarantees. In the case of regular grids we show that the expected approximation error can be bounded (for certain combinations of cost exponent and dimension) independently of the size of the original support of the measures.

Numerical experiments demonstrate the practical performance of the scheme.

Statistical network analysis: From combinatorics to nonparametrics

Sofia Olhede
(joint work with Patrick Wolfe)

Networks or graphs are discrete objects, representing relationships between entities (vertices) in terms of a list of edges. The simplest form of graph has unweighted edges, and does not allow for directed edges or self-loops. It is typical to model each such edge as a Bernoulli random variable taking the value zero or unity.

It is standard practice to collect all these edges in a binary symmetric array, or adjacency matrix. This array will then contain $\binom{n}{2}$ independent random variables, which can be modeled using up to $\binom{n}{2}$ parameters. To understand simplified behavior in this array, as the size of the network grows, the concept of a graph limit has recently emerged (see, *e.g.*, the discussion in [4].) Any infinite binary array that is exchangeable, in the sense that its probability distribution is invariant under symmetric permutations of its indices, can be represented as a realization of a graph limit function, as established by the Aldous–Hoover theorem [3].

Once a network is viewed as a realization from an infinite-dimensional limit object, the question of recovering this limit object becomes a question of nonparametric statistics [5]. One possible choice of estimator relates to the stochastic blockmodel, which implies that certain groups of vertices share similar connectivity properties. Different methods of estimating the parameters of stochastic blockmodels have been proposed [1, 2], and can be shown to imply a convergent approximation of the underlying limit object [5].

Understanding the behavior of a network at the level of groups of vertices is crucial but does not reveal the entire picture. For example, it is possible to define quantities analogous to moments of random variables in the setting of graph limits [1, 2, 3]. We discuss other such representations and their uses in nonparametric statistical methods for the analysis of large graphs.

References

[1] P. J. Bickel, A. Chen, *A nonparametric view of network models and Newman– Girvan and other modularities*, Proceedings of the National Academy of Sciences of the USA **106** (2009), 21068–21073.

[2] P. J. Bickel, A. Chen, E. Levina, *The method of moments and degree distributions for network models*, The Annals of Statistics **39** (2011), 2280–2301.

[3] P. Diaconis, S. Janson, *Graph limits and exchangeable random graphs*, Rend. Mat. Appl. **28** (2008), 33–61.

[4] L. Lovasz, *Large Networks and Graph Limits*, American Mathematical Society, Providence Rhode Island, (2013).

[5] S. C. Olhede, P. J. Wolfe, *Network histograms and universality of blockmodel approximation*, Proceedings of the National Academy of Sciences of the USA **111** (2014), 14722–14727.

Interpretable models for prediction on network-linked data

Levina Elizaveta

(joint work with Tianxi Li, Ji Zhu)

Advances in data collection and social media have resulted in network data being collected in many applications, recording relational information between units of analysis; for example, information about friendships between adolescents is now frequently available in studies of health-related behaviors. This information is often collected along with more traditional covariates on each unit of analysis; in the adolescent example, these may include variables such as age, gender, race, socioeconomic status, academic achievement, etc. Information on friendships can play an important role through network cohesion, the empirically observed phenomenon of friends behaving similarly. Since cohesion suggests pooling information from neighboring nodes, incorporating the network information is potentially helpful in making predictions.

There is a large body of work extending over decades on predicting a response variable of interest from covariates, via linear or generalized linear models, survival analysis, classification methods, and the like, which typically assume the training samples are independent and do not extend to situations where the samples are connected by a network. In certain specific contexts, regression with network dependent observations has been studied. However, most of these methods either lose the interpretability of the classic regression model or cannot make out-of-sample predictions.

We propose a regression model with random node effects incorporating the network information, and a network-based quadratic penalty on these node effects to encourage similarity between predictions for linked nodes. Our model is as interpretable as the classic regression model, in that the network information is modeled by individual effects, decoupled from covariates effects. We show that the method gives consistent estimates of covariate effects and derive explicit conditions on when enforcing network cohesion in regression can be expected to perform better than ordinary least squares. More importantly, our proposal can be directly extended to generalized linear models and survival analysis without conceptual difficulties, as well as to the case of high-dimensional predictors. In our framework, out-of-sample prediction can be easily made and good prediction performance is demonstrated through multiple numerical studies. In contrast to previous work, we assume no specific form for the cohesion effects and require no information about potential groups. We also derive a computationally efficient algorithm for implementing our approach, which is efficient for both sparse and dense networks, the latter with an extra sparsification step which we prove preserves the relevant network properties. To the best of our knowledge, this is the first proposal of a general regression framework with network cohesion among the observations that is computationally feasible and can retain covariate interpretation as well as make out-of-sample predictions.

We apply the proposed method to predict levels of recreational activity and marijuana usage among teenagers based on both demographic covariates and their friendship networks using the data from AddHealth, a national longitudinal study of U.S. high schools. The superior performance of our method compared to standard methods without incorporating network information indicates that social peer effects are indeed very important in these types of applications. In particular, we show that even using the social network information alone can make better predictions than using all the most informative covariates. Our method also outperforms previously proposed ad hoc approaches to incorporating network information. This demonstrates the effectiveness of our method and more generally the practical importance of using network information in prediction problems.

References

[1] Tianxi Li, Elizaveta Levina, and Ji Zhu. *Prediction models for network-linked data*, arXiv preprint arXiv:1602.01192, 2016.

Markov Properties and Geometry of Exchangeable Random Networks

Alessandro Rinaldo

(joint work with Steffen L.. Lauritzen, Kayvan Sadeghi)

We investigate the connections among random network models, bidirected and undirected graphical models, and exchangeability. We show that exchangeable finite network models can be well approximated by mixtures of curved exponential families corresponding to a distinguished class of graphical models for marginal independence of binary data. We obtain a simple derivation of de-Finetti theorem for exchangeable arrays, and we link it to the theory of graphons. Using this characterization, we discuss some of the challenges and intrinsic difficulties of fitting exchangeable network models.

Model-free Knockoffs: Statistical Tools for Replicable Selections

Emmanuel Candès

(joint work with Yingying Fan, Lucas Janson, Jinchi Lv)

Dramatic changes in data acquisition and sharing capabilities have informed a new way of carrying out scientific investigation. Nowadays we routinely collect information on an exhaustive collection of possible explanatory variables to predict an outcome or understand what determines an outcome. For instance, certain diseases have a genetic basis and an important biological problem is to find which genetic features (e.g., gene expressions or single nucleotide polymorphisms) are important for determining a given disease. Even though we believe that a disease status depends on a comparably small set of genetic variations, we have a priori no idea about which ones are relevant and therefore must include them all in our search. In statistical terms, we have an outcome variable Y and a potentially gigantic collection of explanatory variables $X_1, \ldots, X_p$: we would like to know which of the many variables Y depends on. In fact, we would like to do this while controlling a type-I error so that *the results of our investigation do not run into the problem of irreproducibility.*

This talk introduces model-free knockoffs, a framework for finding dependent variables while provably controlling the false discovery rate in finite samples. This feat holds no matter the form of the dependence between Y and X, which does not need to be specified in any way. What is required is that we observe i.i.d. samples and know something about the distribution of the covariates although we have shown that the method is robust to unknown/estimated covariate distributions. This framework builds on the knockoff filter of Foygel Barber and Candès introduced a couple of years ago, which was limited to linear models with fewer variables than observations ($p < n$). Having said this, model-free knockoffs deal with a range of problems far beyond the scope of the original knockoff paper—e.g. it provides valid selections in any generalized linear model including logistic regression—while being more powerful than the original procedure when it applies.

We present an analysis of a genome-wide association study (GWAS) data set with thousands of subjects and hundreds of thousands of single nucleotide polymorphisms (SNP) locations from a case-control study of Crohn's disease in the United Kingdom. Here, model-free knockoffs made twice as many discoveries as the original analysis of the same data. A literature review provides independent confirmation of many of the new discoveries.

References

[1] R. F. Barber and E. J. Candès, Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5) (2015), 2055–2085.

[2] E. J. Candès, Y. Fan, L. Janson and J. Lv, Panning for gold: model-free knockoffs for high-dimensional controlled variable selection. Stanford Technical Report, (2016).

Goodness-of-fit Testing for (non)-Smooth Densities: Local Minimax Rates

Sivaraman Balakrishnan

(joint work with Larry Wasserman)

We consider the classical one-sample goodness-of-fit testing problem of testing a simple null hypothesis that the data are drawn from a specified distribution function, against a composite alternative separated in the total variation metric. We consider testing a Lipschitz density, with possibly unbounded support, in the low-smoothness regime where the Lipschitz parameter is not assumed to be constant. In contrast to classical results, we observe that the minimax rate and critical testing radius in these settings depend strongly and in a precise fashion on the null distribution being tested. For multinomials this phenomenon was recently explicated in the work of [1]. In the full version of this paper we re-visit and extend their results by developing two novel modifications to the χ^2-test whose performance we characterize. In this short abstract we focus on the question of testing Lipschitz densities, we observe that classical binning tests are inadequate in the low-smoothness regime and we design a spatially adaptive partitioning scheme that forms the basis for our locally minimax optimal tests. Furthermore, we provide the first local minimax lower bounds for this problem which yield a precise characterization of the dependence of the critical radius on the null hypothesis being tested.

1. Background and Problem Set-up

We begin with some basic background on hypothesis testing, the testing risk and minimax rates. Our focus in this paper is on the one sample goodness-of-fit testing problem. We observe samples $Z_1, \ldots, Z_n \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^d$, which are independent and identically distributed with distribution P. In this context, for a fixed distribution P_0, we want to test the hypotheses:

$$\begin{aligned} H_0 &: P = P_0 \quad \text{versus} \\ H_1 &: \mathrm{TV}(P, P_0) \geq \epsilon. \end{aligned} \tag{1}$$

It is well-understood [3, 5] that without further restrictions there are no uniformly consistent tests for distinguishing these hypotheses. In the full version of this paper we consider restricted variants of this problem that correspond to testing multinomials and Lipschitz densities. In this abstract we focus on the latter problem.

Minimally smooth density testing: In the density testing problem the set $\mathcal{X} \subset \mathbb{R}^d$, and we restrict our attention to distributions with Lipschitz densities, i.e. letting p_0 and p denote the densities of P_0 and P with respect to the Lebesgue measure, we consider the set of densities:

$$\mathcal{L} = \Big\{ p : \int_{\mathcal{X}} p(x) dx = 1, p(x) \geq 0 \ \forall \ x, |p(x) - p(y)| \leq L_n \|x - y\|_2 \ \forall \ x, y \in \mathbb{R}^d \Big\},$$

and suppose that $p_0, p \in \mathcal{L}$. We particularly emphasize, that unlike prior work [4, 2] we do not restrict the domain of the densities and are interested in the low-smoothness regime where the Lipschitz parameter L_n is allowed to grow with the sample size.

Hypothesis testing and risk: Returning to the setting described in Equation (1), we define a test ϕ as a Borel measurable map, $\phi : \mathcal{X}^n \mapsto \{0, 1\}$. For a fixed null distribution P_0, we define the worst-case risk of the test over a restricted class $\mathcal{C}$ which contains P_0 as:

$$R_n(\phi; P_0, \epsilon, \mathcal{C}) = \mathbb{E}_{P_0}[\phi] + \sup \Big\{ \mathbb{E}_P[1 - \phi] : \|P - P_0\|_1 \geq \epsilon, P \in \mathcal{C} \Big\},$$

where the first term corresponds to the Type-I error and the second term corresponds to the maximum Type-II error. With this definition in place, we can define the minimax risk as,

$$R_n(P_0, \epsilon, \mathcal{C}) = \inf_{\phi} R_n(\phi; P_0, \epsilon, \mathcal{C}). \tag{2}$$

We refer to this quantity as the *local* minimax risk, to emphasize its dependence on P_0. In this paper we view the minimax risk via a coarse lens, focusing on the critical radius or the minimax separation, i.e. the smallest value ϵ for which a hypothesis test has non-trivial power to distinguish P_0 from the set of alternatives. Formally, we define the critical radius as:

$$\epsilon_n(P_0, \mathcal{C}) = \inf \Big\{ \epsilon : R_n(P_0, \epsilon, \mathcal{C}) \leq 1/2 \Big\},$$

where the constant 1/2 is chosen as an arbitrary constant smaller than 1.

2. Main Results

Our main results are briefly stated in this section. We refer the interested reader to the full version of this paper for detailed statements of the results, a development of their consequences and for complete proofs. For a density p_0 we define its bulk $\mathcal{B}_\epsilon$ as the set of smallest Lebesgue measure that contains $1-\epsilon$ probability content. For

$$\gamma = \frac{2}{3+d},$$

we define the truncated γ-norm:

$$\|p_0\|_{\gamma,\epsilon} = \left(\int_{\mathcal{B}_\epsilon} p_0(x)^\gamma \right)^{1/\gamma}.$$

Theorem: [Informal] Let ℓ_n and u_n be defined as solutions to the equations:

$$\ell_n^{4+d} = \frac{L}{n^2} \|p_0\|^2_{\gamma,\ell_n}, \; u_n^{4+d} = \frac{L}{n^2} \|p_0\|^2_{\gamma,u_n/8},$$

then for universal constants $c, C > 0$ we have that the critical radius for the Lipschitz testing problem is bounded as:

$$\ell_n \leq \epsilon_n(p_0, \mathcal{L}) \leq u_n.$$

The upper and lower bounds are based on an adaptive partitioning scheme that divides the domain of the null distribution into bins of a precise width. We defer these technical aspects to the full version of this paper.

A point of emphasis that we conclude with is that this result describes in a fairly precise manner the variation of the critical radius as a function of the null distribution p_0. In particular, it is worth noting that this rate exhibits considerable variability over $\mathcal{L}$, yielding slow rates for testing heavy tailed distributions with large effective support and fast rates for testing spiky null distributions with small effective support.

References

[1] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2014.

[2] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension, 2016.

[3] Andrew R. Barron. Uniformly powerful goodness of fit tests. *Ann. Statist.*, 17(1):107–124, 03 1989.

[4] Yu. I. Ingster. Minimax detection of a signal in ℓ_p metrics. *Journal of Mathematical Sciences*, 68(4):503–515, 1994.

[5] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1(1): 38–53, 01 1973.

Extracting multiscale geometric information from high-dimensional and infinite-dimensional data with application to classification

Wolfgang Polonik
(joint work with Gabriel Chandler)

1. Introduction

We present a method for geometric feature extraction. The extracted features are useful for visualizing geometric aspects of high-dimensional data sets, and for conducting statistical inference such as classification. Our method exhibits multiscale characteristics and is applicable to various types of data, as long as the data live in a Hilbert space. This includes high-dimensional Euclidean data, but also data to which the kernel trick can be applied.

The proposed method is based on the notion of Tukey (halfspace) depth, and has relations to various other methodology known from the literature, including local depth (e.g. see [1, 2, 4]), the shorth-plot ([5]), Choquet capacities, multidimensional scaling, and the concept of mass estimation discussed in ([6]). The latter work also inspired the developments presented here.

2. Constructions of the feature functions

Suppose we are given data $X_1, \ldots, X_n \in \mathbb{R}^d$. For each pair (X_i, X_j), we construct a real-valued function $\widehat{q}_{ij}(\alpha), 0 < \alpha < 1$, which is a quantile function of a certain depth distribution. The collection of $\binom{n}{2}$ functions is then used for statistical analysis of the data set. The depth distribution and the corresponding quantile functions $\widehat{q}_{ij}(\alpha)$ are as follows.

For a given pair of data $(X_i, X_j) \in \mathbb{R}^d \times \mathbb{R}^d$, consider the line $\ell_{ij} = \{s \in \mathbb{R}^d : s = \gamma X_i + (1-\gamma)X_j, \gamma \in \mathbb{R}\}$, and the midpoint $m_{ij} = \frac{1}{2}(X_i + X_j)$. For $s \in \ell_{ij}$ and $\alpha \in (0, \pi)$, let $C_{ij}(s)$ denote the cone with tip s and opening angle α containing m_{ij}. Then $\widehat{d}_{ij}(s)$ is the Tukey depth of m_{ij} among all the data on ℓ_{ij} obtained by projecting all the X_j lying in $C_{ij}(s)$ onto ℓ_{ij}. Formally,

$$\widehat{d}_{ij}(s) = \frac{1}{n} \min \Big\{ \big|\{k : \langle X_k, \tfrac{m_{ij}}{\|m_{ij}\|} \rangle \le \|m_{ij}\| \big|, \ \big|\{k : \langle X_k, \tfrac{m_{ij}}{\|m_{ij}\|} \rangle \ge \|m_{ij}\| \big| \Big\}.$$

We then pick the tip s randomly, independently of the data, and according to a distribution G. This results in a random variable $\widehat{d}_{ij}(S)$, and $\widehat{q}_{ij}(\alpha)$ is defined as the quantile function corresponding to this random variable. Figure 1 shows one set of such functions for the 13-dimensional wine data set (classes 1 and 2 only) from the UC Irvine Machine Learning Repository `http://archive.ics.uci.edu/ml/`. To illustrate the usefulness of these feature functions in classification, the functions $\widehat{q}_{ij}(\alpha)$ are colored according to the class memberships of the points X_i, X_j that are used in constructing the function $q_{ij}(\alpha)$.

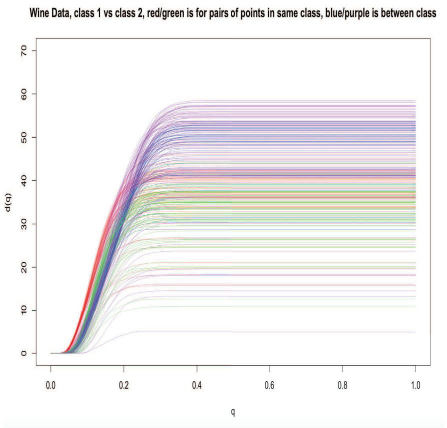

Figure 1. Functions $\widehat{q}_{ij}(\alpha)$ for X_i, X_j running through both class 1 or class 2 of wine data.

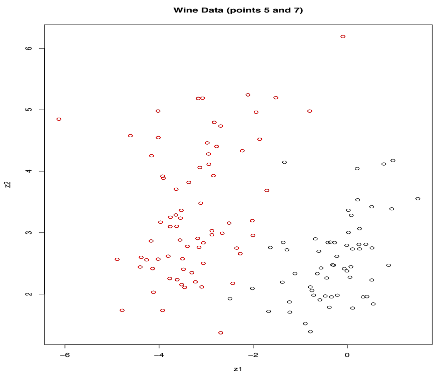

Figure 2. A Z_1-Z_2-plot corresponding to one of the functions on the left panel.

3. Classification

In the case of binary classification, the idea is to use the $\binom{n}{2}$ feature functions to construct, for each data point, a *pair* of (new) feature functions. This results in n pairs of functions, which can be used for classification by utilizing functional data analysis. The construction of the function pairs is as follows. Let $Y_i \in \{0, 1\}$ denote the class label of X_i. For each given X_i, we split up the $\binom{n}{2}$ functions $\widehat{q}_{ij}(\alpha)$ into two subsets, $\mathcal{D}_i^s = \{q_{ij}(\alpha); Y_i = Y_j\}$ and $\mathcal{D}_i^d = \{q_{ij}(\alpha); Y_i \neq Y_j\}$, respectively. The average functions over these subsuts $\widehat{q}_i^s(\alpha) = \text{ave}_{j\in\mathcal{D}_i^s}\widehat{q}_{ij}(\alpha)$, and $\widehat{q}_i^d(\alpha) = \text{ave}_{j\in\mathcal{D}_i^d}\widehat{q}_{ij}(\alpha)$, respectively, comprise the function pairs to be used for classification. A simple approach is to perform functional PCA on the four classes of functions $\{\widehat{q}_i^{(s)}(\alpha),\, Y_i = k\}$, $k = 0, 1$ and $\{\widehat{q}_i^{(d)}(\alpha),\, Y_i = k\}$, $\ell = 0, 1$ separately, and then to combine, for each i, the first k PCA scores from each class into a vector of length $2k$. These vectors can then be used to train an off-the-shelf classifier (such as a kernel SVM), to classify newly incoming unlabelled data. Numerical studies indicate, that this ad-hoc method is competitive with other classification methods.

4. Generalizations and relations to other statistical methods

Shorth plot. The shorth plot is proposed in [5] (see also [3]). It is a concentration measure for one-dimensional functions, geared towards mode finding. For $d = 1$, the function $\widehat{q}_{ij}(\alpha)$ can be shown to be closely related to the shorth plot, but rather than mode finding being the goal, our approach is targeting antimodes.

Local depth. Local depth has been considered in the literature, for instance, in [1, 2, 4]. The approach considered in [4] is perhaps the one closest to our approach, although there are serval methodological differences.

Choquet capacities. A closer investigation of the construction of the average depth quantile functions shows that, for each α, they estimate the expected value of a hitting function of a random closed set. If the data consist of two classes, the functions $\widehat{q}_i^d(\alpha)$ and and $\widehat{q}_i^s(\alpha)$ estimate expectations of two different random sets, the distribution of which depend on whether X_i is compared to points within the same class, or to points in a different class.

Multidimensional scaling. Given a line $\ell \subset \mathbb{R}^d$, depth quantile functions only depend on the number of points in cones (with axis of symmetry being ℓ. To determine this number, all we need are two one-dimensional quantities (depending on line), the (signed) length of projection of a given data point X_k onto ℓ (Z_{k1}, say), and the length of its projection, Z_{k2}. Plotting the pairs (Z_{k1}, Z_{k2}) can be interpreted as a multidimensional scaling method. Note, however, that, each pair (X_i, X_j) results in a line $\ell = \ell_{ij}$, which results in a total of $\binom{n}{2}$ such plots. The functions $\widehat{q}_{ij}(\alpha)$ can be considered as summaries of these plots (see Figure 2).

Non-Euclidean Data. Our method also makes sense in a general Hilbert space setting. One can thus consider non-Euclidean data structures to which the kernel-trick can be applied (functions, networks, etc.), and apply our method to the elements in the corresponding RKHS. This enables an investigation of the corresponding RKHS geometry.

5. Some theoretical results

Let $X_1, \ldots, X_n \sim_{iid} F$, and suppose that both F and G have positive Lebesgue densities. Let $d_{ij}(s)$ and $q_{ij}(\alpha)$ be the theoretical counterparts to $\widehat{d_{ij}}(s)$ and $\widehat{q}_{ij}(\alpha)$, respectively, meaning that in finding the Tukey depth, the empirical distribution is replaced by the true distribution F. Note, however, that both $d_{ij}(s)$ and $q_{ij}(\alpha)$ are still random quantities, as they still depend on the pair (X_i, X_j).

Theorem 1. *As $n \to \infty$, we have*

$$\max_{1 \le i,j \le n} \sup_{s \in \ell_{ij}} \left|\widehat{q}_{ij}(s) - q_{ij}(s)\right| = O_P\Big(\sqrt{\tfrac{\log n}{n}}\Big).$$

Remark. The convergence holds uniformly in the dimension d.

Multiscale nature of methodology: For a pair (X_i, Y_i), let F_{ij} denote the (one-dimensional) distribution on the line given by (X_i, X_j), obtained by (orthogonally) projecting all the mass of F onto this line. With this notation, we have $\lim_{\alpha \to 1} q_{ij}(\alpha) = \min\big(F_{ij}(m_{ij}), 1 - F_{ij}(m_{ij})\big)$ is the global Tukey depth of m_{ij} for the distribution F_{ij}, and $\lim_{\alpha \to 0} \frac{q_{ij}(\alpha)}{\alpha^d} \to c\frac{f(m_{ij})}{g(m_{ij})}$, where $c > 0$ is known, and g is the known pdf of G. This illustrates the multiscale nature of our feature extraction method, as small values of α contain local information about the midpoint (the density), and large values of α contain global information (depth).

References

[1] Agostinelli, C. and Romanazzi, M. (2011): Local Depth. *J. Statist. Plann. Inference* **141**, 817830.

[2] Dutta, S., Chaudhuri, P. and Ghosh, A.K. (2012): Classification using Localized Spatial Depth with Multiple Localization, Technical report.

[3] Einmahl, J.H.J., Gantner, M. and Sawitzki, G. (2010): The Shorth Plot. *J. Comput. Graph. Statist.* **19**, 62-73.

[4] Peindaveine, D. and van Bever, G. (2013): From Depth to Local Depth: A Focus on Centrality. *J. Amer. Statist. Assoc.* **108**, 1105-1119.

[5] Sawitzki, G (1994): Diagnostic Plots for One-Dimensional Data. In: *Computational Statistics. Papers collected on the Occasion of the 25th Conference on Statistical Computing at Schloss Reisensburg.* (Edited by P.Dirschedl & R.Ostermann) Heidelberg, Physica, pp. 237-258.

[6] Ting, K.M., Zhou, G-T., Liu, F.T. and Tan, S.C. (2013): Mass estimation. *Mach. Learn.* **90**, 127-160.

Non-parametric identification of causal effects in the presence of unobserved variables

Thomas S. Richardson

(joint work with Robin J. Evans, James M. Robins, Ilya Shpitser)

Building on prior work by Robins [4], Spirtes et al. [7], Pearl [2] and Tian [8], we described a complete algorithm for the non-parametric identification of causal effects arising from causal directed acyclic graph (DAG) models containing unobserved variables. This algorithm leads directly to a method of deriving algebraic constraints implied by such models. The nested Markov model [1, 3, 5] is defined to be the set of distributions obeying these constraints.

The non-parametric identification question may be stated more precisely as follows: We are given a joint distribution $p(x_H, x_V)$ over observed and hidden variables, indexed by V and H respectively. Further, this distribution factors according to a DAG $\mathcal{G}$:

$$p(x_H, x_V) = \prod_{t \in V \cup H} p(x_t \,|\, x_{\mathrm{pa}(t)}), \tag{1}$$

where $\mathrm{pa}(t)$ is the set of vertices that are parents of t in the DAG $\mathcal{G}$. We wish to know if there is a functional of the observed distribution $p(x_V)$ that identifies the causal *intervention distribution*:

$$p(x_Y \,|\, \mathrm{do}(x_A)) \equiv \sum_{x_{V \setminus (A \cup Y)}} \sum_{x_H} \prod_{t \in H \cup (V \setminus A)} p(x_t \,|\, x_{\mathrm{pa}(t)}), \tag{2}$$

where A, Y are non-empty disjoint subsets of V.

Building on prior work by Tian [8], we show that $p(x_Y \mid \mathrm{do}_{\mathcal{G}(H\cup V)}(x_A))$ is identified if and only if

$$p(x_Y \mid \mathrm{do}(x_A)) = \sum_{x_{Y^*\setminus Y}} \prod_{i=1}^{k} q_{D_i}(x_{D_i} \,|\, x_{(V\cap \mathrm{pa}(D_i))\setminus D_i}), \tag{3}$$

where $Y^* \subseteq V \setminus A$ consists of Y and all non-endpoint vertices in V lying on directed paths from $a \in A$ to $y \in Y$; further $\{D_1, \dots, D_k\}$ is a partition of Y^*, corresponding to c-components [8] in the DAG formed by cutting edges into A. Finally, each of the kernels $q_{D_i}(\cdot \,|\, \cdot)$ is obtained recursively from $p(x_V)$ via a sequence of transformations:

$$q_i^{(j)}(x_{D_i\cup\{v_{j+1},\dots,v_{p_i}\}} \mid x_{\{v_1,\dots,v_j\}}) \equiv \frac{q_i^{(j-1)}(x_{D_i\cup\{v_j,\dots,v_{p_i}\}} \mid x_{\{v_1,\dots,v_{j-1}\}})}{q_i^{(j-1)}(x_j \mid x_{\mathrm{mb}_{j-1}(v_j)})}, \tag{4}$$

with $q_i^{(0)}(x_V) \equiv p(x_V)$ and $q_{D_i} \equiv q_i^{(p_i)}$. Here $\langle v_1, \dots, v_{p_i}\rangle$ is a 'suitable' ordering of the vertices in $V \setminus D_i$ and $\mathrm{mb}_{j-1}(v_j)$ is the Markov blanket of v_j within its c-component in the DAG $\mathcal{G}_i^{(j-1)}$ in which edges have already been cut into $\{v_1, \dots, v_{j-1}\}$. An ordering of $V \setminus D_i$ is *suitable* if, for $j = 1, \dots, p_i$, there is no vertex in $\{v_{j+1}, \dots, v_{p_i}\}$ that is both a descendant of v_j and in the same c-component as v_j in the graph $\mathcal{G}_i^{(j-1)}$. Results in [6] imply that if for some D_i there is no suitable ordering of $V \setminus D_i$ then $p(x_Y \mid \mathrm{do}(x_A))$ is not identified.

The transformation (4) generalizes the usual operations of marginalization and conditioning. The full paper [3] contains results concerning the preservation of conditional independence under this transformation. The paper also describes the Markov properties of the kernels resulting from such transformations, given a marginal from a DAG with hidden variables. It is proved that the model resulting from these non-parametric restrictions is equivalent to the set of distributions obeying the constraints found by Tian's algorithm [9].

References

[1] R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. arXiv preprint: 1511.06813, 2015.

[2] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–709, 1995.

[3] T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. arXiv preprint:1701.06686, 2017.

[4] J. M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[5] I. Shpitser, R. J. Evans, T. S. Richardson, and J. Robins. An introduction to nested Markov models. *Behaviormetrika*, 41(1):3–39, 2014.

[6] I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Twenty-First National Conference on Artificial Intelligence*, 2006.

[7] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 1993.

[8] J. Tian. *Studies in Causal Reasoning and Learning.* PhD thesis, Department of Computer Science, University of California, Los Angeles, 2002.

[9] J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of UAI-02*, pages 519–527, 2002.

Causality, Invariance and Anchors

Jonas Peters

(joint work with Peter Bühlmann, Christina Heinze-Deml, Nicolai Meinshausen and Niklas Pfister)

In the field of causality we want to understand how a system reacts under interventions (e.g. in gene knock-out experiments). These questions go beyond statistical dependences and can therefore not be answered by standard regression or classification techniques. In order to answer them, one requires a causal model of the underlying system. A causal model entails not only an observational distribution but also intervention distributions, i.e., it predicts the system's behaviour under interventions. In this talk, we use the language of structural causal models (SCMs) that are also sometimes called structural equation models or functional causal models [6]. In SCMs, each variable $X_j, j = 1, \dots, d$ is a deterministic function of its direct causes $X_{pa(j)}$ and some noise variable N_j, that is

$$X_j := f_j(X_{pa(j)}, N_j), \quad j = 1, \dots, d,$$

where all noise variables are assumed to be jointly independent. The corresponding graph is obtained when drawing edges from the variables appearing on the right hand side to variables appearing on the left hand side. This graph is assumed to be acyclic, i.e., it does not contain any directed cycle. Due to the acyclicity, it is apparent that an SCM entails a joint distribution $P(X_1, \dots, X_d)$ over the variables $X_1, \dots, X_d$. Intervention distributions are obtained when one of the structural assignments is replaced by the intervened version. For example, for the intervention $do(X_4 := 5)$ we replace the fourth structural assignment with $X_4 := 5$. In causal discovery (or structure learning), we are given an i.i.d. data set from the joint distribution $P(X_1, \dots, X_d)$ and try to learn the causal structure. An overview of ideas and methods can be found in [8], for example. In this work, instead of learning the whole graph, we focus on a special subclass of this problem: We assume that we are given a target variable Y and aim at learning the direct causes of Y, i.e., its causal parents from the set $\{X_1, \dots, X_d\}$ of covariates. In order to do so, we propose to exploit invariances with respect to so-called anchors. We want to illustrate this idea with three examples.

1. Discrete Set of Environments as Anchor

We assume that we are given $(X_1, Y_1), \dots, (X_n, Y_n)$ and a set $\mathcal{E}$ of finitely many environments $\mathcal{E} = \{e_1, \dots, e_k\}$ that describe a decomposition of the sample $\{1, \dots, n\}$. For example, $e_1 = \{1, 2, \dots, 40\}$, $e_2 = \{41, \dots, 100\}$, $e_3 = \{101, \dots, n\}$.

We now make the following invariance assumption: H_{0,X_S} is true for a set $S \subseteq \{1, \ldots, d\}$ if for all $i = 1, \ldots, n$ we have

$$Y_i = X_i \cdot \gamma + \varepsilon_i$$

and furthermore $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d.. Note that for all i, X_i can have an arbitrary distribution.

Both here, and in the two scenarios below, we assume that $H_{0,X_{S^*}}$ is satisfied for a set S^*. For example, if the environments correspond to different intervention distributions, and the interventions act on any variable(s) other than Y, $H_{0,X_{S^*}}$ is true for S^* being the direct parents of Y [7]. This follows from the assumption of modularity, autonomy or stability [3, 1, 5, 6, 10].

The key idea is to test H_{0,X_S} for different sets S at level α (see [7] for such tests) and then define $\hat{S}$ as follows:

$$\hat{S} := \bigcap_{S: H_{0,X_S} \text{ not rejected}} S, \tag{1}$$

and $\hat{S} := \emptyset$ if the index set is empty. This implies the guarantee

$$P(\hat{S} \subseteq S^*) \geq 1 - \alpha. \tag{2}$$

The environments provide us with power in order to reject wrong sets $S \neq S^*$.

2. Non-descendant Variable as Anchor

If we are not given environments, we may use one of the covariates as an anchor. More formally, we assume we are given data $(X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)$, where, again, X denotes the d-dimensional covariates and Y denotes the target variable. We now define that a subset $S \subseteq \{1, \ldots, d\}$ of covariates satisfies the invariance property H_{0,X_S} if and only if $(X_1, Y_1, E_1), \ldots, (X_n, Y_n, E_n)$ are i.i.d. and

$$E \perp\!\!\!\perp Y \,|\, X,$$

where $\perp\!\!\!\perp$ denotes statistical independence. If the joint distribution over (X, E, Y) is induced by an SCM, then $H_{0,X_{S^*}}$ is true as long as E is a non-descendant of Y. Again, the estimator $\hat{S}$ can be defined as in (1) and we obtain the guarantee (2).

The component that is still missing is a testing procedure for H_{0,X_S}. One may use a kernel-based conditional independence test [11], for example. Alternatively, one may a priori restrict the dependence model and test whether E is significant in a regression model $Y \sim X, E$, for example (or, conversely, whether Y is significant in a regression $E \sim X, Y$). [4] compare these and more testing procedures and apply the methodology to a real world data set.

3. Time as Anchor

Lastly, we may also use a time index as an anchor. In this setup, we assume we are given $(X_1, Y_1), \ldots, (X_n, Y_n)$ and think about obtaining the data points one after each other. We now define that a subset $S \subseteq \{1, \ldots, d\}$ of covariates satisfies the invariance assumption H_{0,X_S} if and only if for all $t = 1, \dots, n$:

$$Y_t = X_t \cdot \gamma + \varepsilon_t$$

and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d.. Note that, again, X_t can have an arbitrary distribution. As before, we can relate this condition to causality: If the structural assignments change over time (but not the one for Y), then, $H_{0,X_{S^*}}$ is satisfied for S^* being the set of direct parents of Y. As before, the estimator $\hat{S}$ can be defined as in (1) and we obtain the guarantee (2). This time, the testing procedure requires some thoughts. E.g., one may put a grid on the time axis that decomposes the time indices $\{1, \dots, n\}$ into ten segments, say. One can then construct environments by joining neighbouring segments (which creates environments of the form $\{k, k+1, \dots, \ell - 1, \ell\}$) and then test whether every pair of environments allows for the same regression model from Y on X_S. This test can be performed by a test statistic based on [2], for example. [9] provide more details including consistency results and applications to real data.

References

[1] J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

[2] G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[3] T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

[4] C. Heinze-Deml, J. Peters, and N. Meinshausen. Predicting the effect of interventions using invariance principles for nonlinear models. NIPS Workshop 2016, 2016.

[5] K. D. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

[7] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (with discussion)*, 78(5):947–1012, 2016.

[8] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. forthcoming, already available online: `http://www.math.ku.dk/~peters/`.

[9] N. Pfister, P. Bühlmann, N. Meinshausen, and J. Peters. Invariant causal prediction for sequential data. *(in preparation)*, 2017.

[10] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

[11] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.

Eigenvectors of Orthogonally Decomposable Functions

Misha Belkin

(joint work with Luis Rademacher, James Voss)

Eigendecomposition of quadratic forms guaranteed by the spectral theorem is the foundation for many important algorithms in computer science, data analysis, and machine learning. In this talk I will discuss our recent work on generalizations from quadratic forms to a broad class of functions based on an analogue of the spectral decomposition in an orthogonal basis. We call such functions "orthogonally decomposable". It turns out that many inferential problems of recent interest including orthogonal tensor decompositions, Independent Component Analysis (ICA), topic models, spectral clustering, and Gaussian mixture learning can be viewed as recovering basis elements from non-quadratic functions of this type.

We identify a key role of convexity in extending traditional characterizations of eigenvectors to the more generic setting of orthogonally decomposable functions. We focus on extending two traditional characterizations of eigenvectors: First, that the eigenvectors of a quadratic form arise from the optima structure of the quadratic form on the sphere, and second that the eigenvectors are the fixed points of the power iteration. Our generalization of the power iteration is a simple first order algorithm, "gradient iteration".

This gradient iteration leads to efficient and easily implementable methods for basis recovery, including such methods as cumulant-based FastICA and the tensor power iteration for orthogonally decomposable tensors as special cases. I will discuss our theoretical analysis of gradient iteration using the structure theory of discrete dynamical systems to show almost sure convergence and fast (super-linear) convergence rates.

The analysis is extended to the case when the observed function is only approximately orthogonally decomposable, with bounds that are polynomial in dimension and other relevant parameters, such as perturbation size. The results can be considered as a non-linear version of the classical Davis-Kahan theorem for perturbations of eigenvectors of symmetric matrices.

Finally, I will go over some new applications of the proposed framework to spectral clustering.

References

[1] M. Belkin, L. Rademacher, J. Voss. *Learning a Hidden Basis Through Imperfect Measurements: An Algorithmic Primitive*, Computational Learning Theory (COLT), 2016.

[2] M. Belkin, L. Rademacher, J. Voss. , *The Hidden Convexity of Spectral Clustering*, AAAI-16: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

Geodesic convexity and regularized scatter estimation

Lutz Duembgen

(joint work with David E. Tyler, Klaus Nordhausen, Heike Schuhmacher)

As noted by [1] and [6], for a thorough understanding of estimation of covariance matrices it may be helpful to view the space of symmetric, positive definite matrices ('scatter matrices') Σ of size $q \times q$ as a Riemannian manifold with local inner product

$$\langle A, B \rangle_\Sigma \ := \ \mathrm{trace}(A\Sigma^{-1}B\Sigma^{-1})$$

of two symmetric matrices A, B of size $q \times q$. This may be motivated by considering Wishart distributions. The resulting geodesic distance between two scatter matrices Σ_0, Σ_1 is given by

$$D_g(\Sigma_0, \Sigma_1) \ := \ \big\|\log(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\big\|_F$$

with $\|\cdot\|_F$ denoting the Frobenius norm, see [2]. These considerations show that for a given scatter matrix $\Sigma = BB^\top$, a natural local parametrization of arbitrary scatter matrices is given by $B\exp(A)B^\top$, where A is running through all symmetric matrices of size $q \times q$, see [3].

Related to this concept is the notion of geodesic convexity (g-convexity): A function $f(\Sigma)$ is called (strictly) geodesically convex if $f(B\exp(tA)B^\top)$ is a (strictly) convex function of $t \in \mathbb{R}$ for any choice of $\Sigma = BB^\top$ and $A \neq 0$. It turns out that the target functions (minus log-likelihood) underlying multivariate M-functionals of scatter are typically g-convex. Under mild regularity conditions they are even strictly g-convex and geodesically coercive (g-coercive) in the sense that

$$f(\Sigma) \ \to \ \infty \quad \text{as } D_g(I, \Sigma) = \|\log(\Sigma)\|_F \to \infty.$$

see [5].

In high-dimensional settings, however, strict g-convexity or g-coercivity are no longer satisfied unless one is using some regularization. So instead of minimizing a g-convex function $L(\Sigma, P)$ depending on a (true or empirical) distribution P on $\mathbb{R}^q$ one tries to minimize

$$f(\Sigma) \ = \ L(\Sigma, P) + \alpha \mathrm{Pen}(\Sigma)$$

for some tuning parameter $\alpha > 0$ and a g-convex penalty function $\mathrm{Pen}(\Sigma)$. Examples for such penalties are

$$\begin{aligned}
\mathrm{Pen}_0(\Sigma) &= \log\mathrm{trace}(\Sigma) + \log\mathrm{trace}(\Sigma^{-1}),\\
\mathrm{Pen}_1(\Sigma) &= q^{-1}\log\det(\Sigma) + \log\mathrm{trace}(\Sigma^{-1}),\\
\mathrm{Pen}_2(\Sigma) &= \log\det(\Sigma) + q\log\lambda_{\max}(\Sigma),\\
\mathrm{Pen}_3(\Sigma) &= \log\big(\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)\big).
\end{aligned}$$

As shown by [5], (strict) g-convexity and g-coercivity of such penalized functions $f(\Sigma)$ may be verified rather easily. The explicit computation of a minimizer is feasible via a partial Newton algorithm similar to the one of [4], and the tuning parameter α may be chosen by a cross validation scheme.

References

[1] C. Auderset, C. Mazza and E. A. Ruh, *Angular Gaussian and Cauchy estimation*, J. Multivar. Anal. **93** (2005), 180–197.

[2] R. Bhatia, *Positive definite matrices*, Princeton University Press (2007)

[3] L. Dümbgen, M. Pauly and T. Schweizer, *M-functionals of multivariate scatter*, Stat. Surv. **9** (2015), 32–105.

[4] L. Dümbgen, K. Nordhausen and H. Schuhmacher, *New algorithms for M-estimation of multivariate scatter and location*, J. Multivar. Anal. **144** (2016), 200–217.

[5] L. Dümbgen and D.E. Tyler, *Geodesic convexity and regularized scatter estimators*, Preprint (arxiv 1607.05455)

[6] A. Wiesel, *Geodesic convexity and covariance estimation*, IEEE Trans. Signal Process. **60** (2012), 6182–6189.

Efficient multivariate entropy estimation via k-nearest neighbour distances

Richard J. Samworth

(joint work with Thomas B. Berrett, Ming Yuan)

Many statistical procedures, including goodness-of-fit tests and methods for independent component analysis, rely critically on the estimation of the entropy of a distribution. In this work, we seek entropy estimators that are efficient in the sense of achieving the local asymptotic minimax lower bound. To this end, we initially study a generalisation of the estimator originally proposed by [2], based on the k-nearest neighbour distances of a sample of n independent and identically distributed random vectors in $\mathbb{R}^d$. When $d \leq 3$ and provided $k/\log^5 n \to \infty$ (as well as other regularity conditions), we show that the estimator is efficient; on the other hand, when $d \geq 4$, a non-trivial bias precludes its efficiency regardless of the choice of k. This motivates us to consider a new entropy estimator, formed as a weighted average of Kozachenko–Leonenko estimators for different values of k. A careful choice of weights enables us to obtain an efficient estimator in arbitrary dimensions, given sufficient smoothness.

References

[1] T. B. Berrett, R. J. Samworth and M. Yuan, *Efficient multivariate entropy estimation via k-nearest neighbour distances*, (2016) `https://arxiv.org/abs/1606.00304`.

[2] L. F. Kozachenko and N. N. Leonenko, *Sample estimate of the entropy of a random vector*, Probl. Inform. Transm., **23** (1987), 95–101.

Group invariance and computational sufficiency for regularized M-estimators

Vincent Q. Vu

The estimation of high-dimensional matrices arises naturally in multivariate problems involving the inference of pairwise relationships between many variables (or entities) based on limited samples. Examples include precision matrices, covariance matrices, MRFs, and PCA. Many estimators proposed for these problems are based on penalized likelihood or loss, because some form of regularization is usually necessary to ensure good statistical properties. However, the computation of these estimators may not scale well with the size of the problem—typically cubic or worse time complexity. We show that in a large class of such problems, the efficient computation of these estimators can be enabled by symmetries of the problem and sufficient regularization.

Adaptive nonparametric Clustering

Vladimir Spokoiny
(joint work with Kirill Efimov, Larisa Adamyan)

This paper aims at offering a novel approach to the classical problem of nonparametric clustering using the idea of multiscale testing of the “no gap” hypothesis. This idea differs significantly from the usual density based approach and allows to distinguish overlapping or connected clusters as well as manifold structures. The resulting procedure called Adaptive weights Clustering (AWC) is fully adaptive and does not require to specify the number of clusters. The clustering results are not sensitive to noise and outliers, the procedure is able to recover different clusters with sharp edges or manifold structure. Our intensive numerical study shows a state-of-the-art performance of the method for a wide range of artificial and real life examples.

Let $\{X_1, \ldots, X_n\}$ be the set of all samples $X_i \in \mathbb{R}^p$. Here the dimension p can be very large or even infinite. The proposed procedure operates with the distance (or similarity) matrix $\bigl(d(X_i, X_j)\bigr)_{i,j=1}^{n}$ only. For every point X_i, the clustering procedure attempts to describe its largest possible local neighborhood in which the data is homogeneous in a sense of a spatial data separation. The clustering structure of the data is described in terms of binary weights w_{ij}, where $w_{ij} = 1$ indicates being points X_i and X_j in the same cluster, whereas $w_{ij} = 0$ means that these points belong to different clusters. Thus, the whole clustering structure of the data can be described using the matrix of weights.

The proposed procedure is sequential and attempts to recover the weights w_{ij} from the data, which explains the name “adaptive weights clustering”. It starts with very local clustering structure $C_i^{(0)}$, that is, the starting positive weights $w_{ij}^{(0)}$ are limited to the closest neighbors X_j of the point X_i in terms of the distance $d(X_i, X_j)$. At each step $k \geq 1$, the weights $w_{ij}^{(k)}$ are recomputed by means of statistical tests of “no gap” between $C_i^{(k-1)}$ and $C_j^{(k-1)}$, the local clusters on step $k-1$ for points X_i and X_j correspondingly.

First of all we fix a growing sequence of radii $h_1 \leq h_2 \leq \ldots \leq h_K$ which determines how fast the algorithm will come from considering very local structures to large-scale objects. Each value h_k can be viewed as a resolution (scale) of the method at step k. The rule has to ensure that the average number of screened neighbors for each X_i at step k grows at most exponentially with $k \geq 1$.

Suppose that the first $k-1$ steps of the iterative procedure have been carried out. The resulting collection of weights $\bigl\{w_{ij}^{(k-1)}, \, j = 1, \ldots, n\bigr\}$ for each point X_i describe a local “cluster” associated with X_i. At the next step k we pick up a larger radius h_k and recompute the weights $w_{ij}^{(k)}$ using the previous results. Only points with $d(X_i, X_j) \leq h_k$ have to be screened at step k. The basic idea behind

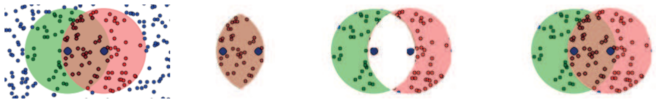

FIGURE 1. From left: Homogeneous case; $N^{(k)}_{i\wedge j}$; $N^{(k)}_{i\triangle j}$; $N^{(k)}_{i\vee j}$

the definition of $w^{(k)}_{ij}$ is to check for each pair i, j with $d(X_i, X_j) \le h_k$ whether the related clusters are well separated or they can be aggregated into one homogeneous region. The test compares the data density in the union and overlap of two clusters associated with the points X_i and X_j. The formal definition involves the weighted empirical mass of the overlap and the weighted empirical mass of the union of two balls $B(X_i, h_{k-1})$ and $B(X_j, h_{k-1})$ shown on Figure 1. *The empirical mass of the overlap* $N^{(k)}_{i\wedge j}$ can be naturally defined as

$$N^{(k)}_{i\wedge j} = \sum_{l\neq i,j} w^{(k-1)}_{il} w^{(k-1)}_{jl} .$$

Similarly, the *mass of the complement* is defined as

$$N^{(k)}_{i\triangle j} = \sum_{l\neq i,j} \bigl\{ w^{(k-1)}_{il} \, \mathbb{1}\bigl(X_l \notin B(X_j, h_{k-1})\bigr) + w^{(k-1)}_{jl} \, \mathbb{1}\bigl(X_l \notin B(X_i, h_{k-1})\bigr) \bigr\}.$$

Note that $N^{(k)}_{i\triangle j}$ is nearly the number of points in $C^{(k-1)}_i$ and $C^{(k-1)}_j$ which do not belong to the overlap $B(X_i, h_{k-1}) \cap B(X_j, h_{k-1})$. Finally, *mass of the union* $N^{(k)}_{i\vee j}$ can be defined as the sum of the mass of overlap and the mass of the complement: $N^{(k)}_{i\vee j} = N^{(k)}_{i\wedge j} + N^{(k)}_{i\triangle j}$. To measure the gap, consider the ratio of these two masses:

$$\tilde{\theta}^{(k)}_{ij} = N^{(k)}_{i\wedge j} \, / \, N^{(k)}_{i\vee j} \, .$$

This value can be viewed as an estimate of the value θ_{ij} which measures the ratio of the averaged density in the overlap of two local regions $C^{(k-1)}_i$ and $C^{(k-1)}_j$ relative to the average density. Under local homogeneity one can suppose that the density in the union of two balls is nearly constant. In this case, the estimate $\tilde{\theta}^{(k)}_{ij}$ should be close to the ratio of the volume of overlap and the volume of union of these balls:

$$\tilde{\theta}^{(k)}_{ij} \approx q^{(k)}_{ij} = \frac{V_\cap(d_{ij}, h_{k-1})}{2V(h_{k-1}) - V_\cap(d_{ij}, h_{k-1})},$$

where $V(h)$ is the volume of a ball with radius h and $V_\cap(d, h)$ is the volume of the intersection of two balls with radius h and the distance $d_{ij} = d(X_i, X_j)$ between centers. The new value $w^{(k)}_{ij}$ can be viewed as a randomized test of the

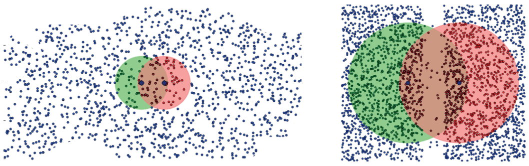

Figure 2. Left: Homogeneous case. Right: “Gap” case.

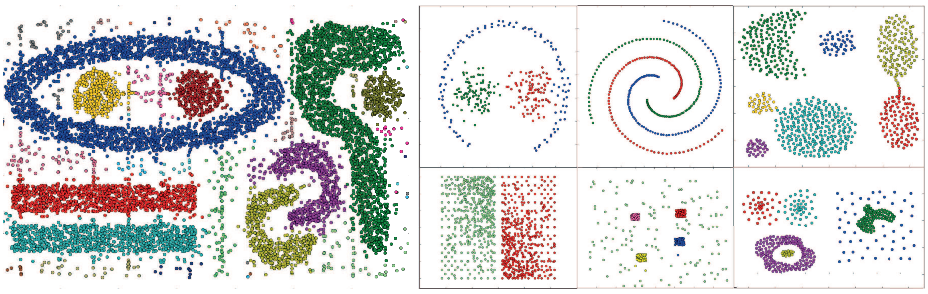

Figure 3. AWC performance on artificial datasets.

null hypothesis $H_{ij}^{(k)}$ of no gap between X_i and X_j against the alternative of a significant gap; see Figure 2 for two examples without and with a gap. The gap is significant if $\tilde{\theta}_{ij}^{(k)}$ is significantly smaller than $q_{ij}^{(k)}$. The likelihood ratio test satistic of “no gap” between two local clusters reads as

$$T_{ij}^{(k)} = N_{i\vee j}^{(k)}\, KL\big(\tilde{\theta}_{ij}^{(k)}, q_{ij}^{(k)}\big) \left\{ \mathbb{I}(\tilde{\theta}_{ij}^{(k)} \le q_{ij}^{(k)}) - \mathbb{I}(\tilde{\theta}_{ij}^{(k)} > q_{ij}^{(k)}) \right\},$$

where $KL(\theta, \eta)$ is the Kullback-Leibler (KL) divergence between two Bernoulli laws. Due to the famous Wilks result this likelihood ratio test statistic is nearly χ^2 which helps to measure the significance level in a unified manner using the threshold λ.

The weights $w_{ij}^{(k)}$ are updated for all pairs X_i, X_j with distance $d(X_i, X_j) \le h_k$:

$$w_{ij}^{(k)} = \mathbb{I}\big(d(X_i, X_j) \le h_k\big)\ \mathbb{I}\left(T_{ij}^{(k)} \le \lambda\right)$$

The next theorem states optimality of the proposed procedure for the case of two close clusters separated by a hole with a lower density.

Theorem 1. *1) Let the data support V contain a fixed hole G, and the data density $f(\cdot)$ be equal to f_1 on the complement $V \setminus G$ and to $f_0 = (1 - \varepsilon_N) f_1$ on G. If $N\varepsilon_N^2 \le C, N \to \infty$ for a fixed constant C, then there is no method which can consistently separate the cases with $\varepsilon_N = 0$ (no gap) and $\varepsilon_N > 0$.*

2) Let a set V be split by a hole G with $\delta = |G|/|V| \geq 1/3$ and $f(\cdot)$ fulfill

$$f(x) \leq (1-\varepsilon) f_1, \quad x \in G, \qquad f(x) \geq f_1\,, \quad x \in V \setminus G$$

Let $X_i \in V_1$, $X_j \in V_2$ and $N\varepsilon^2 \geq C \log(N)$ for a fixed sufficiently large constant C. Then $w_{ij}^{(k)} = 0$ with a high probability.

Nonparametric Bayes for an irregular model

Johannes Schmidt-Hieber

(joint work with Markus Reiß)

Suppose we observe a Poisson point process on $[0,1] \times \mathbb{R}$ with intensity $\lambda_f(x,y) = n\mathbf{1}(f(x) \leq y)$. The statistical problem is to recover the support boundary f from the data. If f is constant the support boundary problem is equivalent to observing n i.i.d. copies of $Y = f + \varepsilon$ with $\varepsilon \sim \mathrm{Exp}(1)$. This model is not Hellinger differentiable. Support boundary recovery of a Poisson point process can therefore be viewed as a nonparametric irregular model.

For parametric irregular models, it is well-known that Bayesian methods outperform the maximum likelihood estimator (MLE). Applying nonparametric Bayes procedures in the support recovery model is therefore natural. Under some assumptions on the parameter space the nonparametric MLE exists in this model. This allows us to compare Bayes directly with the likelihood method. For estimation of the functional $\vartheta = \int f$, the MLE is $\widehat{\vartheta}^{\mathrm{MLE}} = \int \widehat{f}^{\mathrm{MLE}}$. This estimator is typically not rate-optimal. For a point process $N = \sum_i \delta_{(X_i,Y_i)}$ on $[0,1] \times \mathbb{R}$ the pairs $(X_i, Y_i)_i$ are called support points. A better estimator for $\vartheta = \int f$ is given by

$$\widehat{\vartheta} = \int \widehat{f}^{\mathrm{MLE}} - \frac{\text{number of support points on the MLE}}{n}.$$

The estimator $\widehat{\vartheta}$ is in many cases rate-optimal and even unbiased with minimal variance (UMVU), cf. [2]. Notice that the MLE and the estimator $\widehat{\vartheta}$ differ by the correction term (number of support points on the MLE)$/n$, which is not easy to motivate. A natural question is to ask whether a Bayesian approach would automatically correct the MLE.

To analyze the posterior distribution we study posterior contraction rates and Bernstein-von Mises theorems. Since the model is irregular, the likelihood ratios only exist under some partial ordering of the support boundary function. It can be shown that under the frequentist distribution P_{f_0} the Bayes formula can be written as

$$\Pi(B|N) = \frac{\int_B e^{-n\int (f_0-f)_+} \frac{dP_{f_0 \vee f}}{dP_{f_0}}(N) d\Pi(f)}{\int e^{-n\int (f_0-f)_+} \frac{dP_{f_0 \vee f}}{dP_{f_0}}(N) d\Pi(f)},$$

almost surely. The L^1-distance is in the support boundary recovery model the intrinsic loss induced by the information geometry. Because of the irregularity of the model, the well-known general meta-theorem for posterior contraction rates in [3] cannot be applied here. Using one-sided analogs of the conditions in the meta-theorem it is, however, possible to prove a modification of the result that is applicable. With this modification we can derive posterior contraction rates for Gaussian priors and (truncated) random series priors. For a class of hyper-priors on the truncation level, it can be shown that the contraction rate matches the adaptive estimation rate up to $\log n$ factors. We also study compound Poisson process priors and show that the posterior also contracts with the adaptive estimation rate for smoothness indices at most one (again up to $\log n$ factors). For monotone support boundaries, we consider subordinator priors and study the dependence of the jump measure on the contraction rates.

While we can derive posterior contraction rates for various families of priors and function classes of candidate support boundaries, results on the limiting shape of the posterior are extremely difficult to establish. We only investigate the case of piecewise constant functions with number of pieces K_n growing to infinity. This allows us already to get some interesting insights into the Bayesian correction of the MLE. For a "nice" class of priors, it is possible to prove that the marginal posterior on $\vartheta = \int f$ converges in the Bernstein-von Mises sense (in total variation under the frequentist distribution) to a $\mathcal{N}(\widehat{\vartheta}, K_n/n^2)$ random variable. The posterior concentrates therefore around the corrected estimator $\widehat{\vartheta}$ and not around the MLE. This property is, however, lost under model misspecification. If the true function is piecewise linear, then the posterior will still correct the MLE but by the wrong amount. In this case there are frequentist estimators that outperform the Bayesian approach.

References

[1] R. Reiß and J. Schmidt-Hieber, *Nonparametric Bayesian analysis for support boundary recovery*, arXiv preprint 1703.08358 (2017).

[2] R. Reiß and L. Selk, *Efficient estimation of functionals in nonparametric boundary models*, Bernoulli **23** (2017), 1022–1055.

[3] S. Ghoshal, J. G. and A. van der Vaart, *Convergence rates of posterior distributions*, Ann. Statist. **28** (2000), 500–531.

Statistical and Computational Guarantees of Lloyd's Algorithm and Its Variants

Harrison H. Zhou

(joint work with Yu Lu)

Lloyd's algorithm, proposed in 1957 by Stuart Lloyd at Bell Labs, is still one of the most popular clustering algorithms used by practitioners, with a wide range of applications from computer vision, astronomy, and to biology. Although considerable innovations have been made on developing new provable and efficient clustering algorithms in the past six decades, Lloyd's algorithm has been consistently listed as one of the top ten data mining algorithms in several recent surveys.

Lloyd's algorithm is very simple and easy to implement. It starts with an initial estimate of centers or labels and then iteratively updates the labels and the centers until convergence. Despite its simplicity and a wide range of successful applications, surprisingly, there is little theoretical analysis on explaining the effectiveness of Lloyd's algorithm. It is well known that there are two issues with Lloyd's algorithm under the worst case analysis. First, as a greedy algorithm, Lloyd's algorithm is only guaranteed to converge to a local minimum. The k-means objective function that Lloyd's algorithm attempts to minimize is NP-hard. Second, the convergence rate of Lloyd's algorithm can be very slow. Arthur and Vassilvitskii construct a worst-case showing that Lloyd's algorithm can require a superpolynomial running time.

A main goal of this paper is trying to bridge this gap between theory and practice of Lloyd's algorithm. We analyze its performance on the Gaussian mixture model, a standard model for clustering, and consider the generalization to sub-Gaussian mixtures, which includes binary observations as a special case. Specifically, we attempt to address following questions to help understand Lloyd's algorithm: How good does the initializer need to be? How fast does the algorithm converge? What separation conditions do we need? What is the clustering error rate and how it is compared with the optimal statistical accuracy?

In this paper, we give a considerably weak initialization condition under which Lloyd's algorithm converges to the optimal label estimators of sub-Gaussian mixture model. While previous results focus on exact recovery (strong consistency) of the labels, we obtain the clustering error rates of Lloyd's algorithm under various signal-to-noise levels. As a special case, we obtain exact recovery with high probability when the signal-to-noise level is bigger than $4 \log n$. The signal-to-noise ratio condition for exact recovery is weaker than the state-of-the-art result. In contrast to previous two-stage (two-step) estimators, our analyses go beyond one-step update. We are able to show a linear convergence to the statistical optimal error rate for Lloyd's algorithms and its two variants for community detection and crowdsourcing.

We illustrate our contributions here by considering the problem of clustering two-component spherical Gaussian mixtures, with symmetric centers θ^* and $-\theta^* \in \mathbb{R}^d$ and variance σ^2. Let n be the sample size and $r = \|\theta^*\|/(\sigma\sqrt{1+9d/n})$ be the normalized signal-to-noise ratio. We establish the following basin of attractions of Lloyd's algorithm.

Theorem. Assume $r \geq C$ and $n \geq C$ for a sufficiently large constant C. For symmetric, two-component spherical Gaussian mixtures, given any initial estimator of labels with clustering error

$$A_0 < \frac{1}{2} - \frac{2.56 + \sqrt{\log r}}{r} - \frac{1}{\sqrt{n}}, \quad w.h.p.$$

Lloyd's algorithm converges linearly to an exponentially small rate after $\lceil 3 \log n \rceil$ iterations, which is the minimax rate as $r \to \infty$ w.h.p.

The results above are extended to general number of clusters k and to (non-spherical) sub-Gaussian distributions under an appropriate initialization condition and a signal-to-noise ratio condition, which, to the best of our knowledge, are the weakest conditions in literature.

References

[1] Yu Lu and Harrison H. Zhou, *Statistical and Computational Guarantees of Lloyd's Algorithm and Its Variants*, http://www.stat.yale.edu/hz68/Lloyd.pdf.

Reporter: Axel Munk

Participants

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

Prof. Dr. Sivaraman Balakrishnan
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
UNITED STATES

Merle Behr
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
37077 Göttingen
GERMANY

Prof. Dr. Misha Belkin
Department of Computer Science
and Engineering
Ohio State University
2015 Neil Avenue
Columbus OH 43210-1277
UNITED STATES

Prof. Dr. Gilles Blanchard
Institut für Mathematik
Universität Potsdam
Karl-Liebknecht-Straße 24-25
14476 Potsdam
GERMANY

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

Prof. Dr. Emmanuel Candes
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

Dr. Marco Cuturi
École Nationale de la Statistique
e de l'Adm. Economique
ENSAE - CREST
3, Avenue Pierre Larousse
92240 Malakoff Cedex
FRANCE

Miguel del Alamo
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
37077 Göttingen
GERMANY

Alexis Derumigny
ENSAE - CREST
Timbre J 340
3, Avenue Pierre Larousse
92240 Malakoff Cedex
FRANCE

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

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

Prof. Dr. Lutz Dümbgen
Institut für Mathematische Statistik
und Versicherungslehre
Universität Bern
Alpeneggstrasse 22
3012 Bern
SWITZERLAND

Raaz Dwivedi
Department of Mathematics
University of California, Berkeley
970 Evans Hall
Berkeley CA 94720-3840
UNITED STATES

Andreas Elsener
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Stephan Huckemann
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Claudia König
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Solt Kovacs
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Elizaveta Levina
Department of Statistics
University of Michigan
311 West Hall
1085 S. University Avenue
Ann Arbor, MI 48109-1107
UNITED STATES

Prof. Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 205
69120 Heidelberg
GERMANY

Prof. Dr. Nicolai Meinshausen
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Sayan Mukherjee
Department of Statistical Science
Duke University
112 Old Chemistry Building
P.O. Box 90251
Durham NC 27710
UNITED STATES

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

Prof. Dr. Long Nguyen
Department of Statistics
The University of Michigan
1447 Mason Hall
Ann Arbor, MI 48109-1027
UNITED STATES

Prof. Dr. Richard Nickl
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Dr. Sofia Olhede
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Prof. Dr. Victor Panaretos
Section de Mathématiques
Station 8
École Polytechnique Fédérale de
Lausanne
1015 Lausanne
SWITZERLAND

Prof. Dr. Jonas Peters
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Niklas Pfister
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

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

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

Prof. Dr. Thomas S. Richardson
Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
UNITED STATES

Prof. Dr. Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139-4307
UNITED STATES

Prof. Dr. Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

Prof. Dr. Angelika Rohde
Fakultät für Mathematik
Albert-Ludwigs-Universität Freiburg
LST für Stochastik
Eckerstrasse 1
79104 Freiburg i. Br.
GERMANY

Dominik Rothenhäusler
Seminar für Statistik
ETH Zürich (HG G 18)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Richard Samworth
Statistical Laboratory
Centre for Mathematical Sciences
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Dr. Johannes Schmidt-Hieber
Mathematical Institute
University of Leiden
Niels Bohrweg 1
2300 RA Leiden
NETHERLANDS

Dr. Rajen Dinesh Shah
Department of Pure Mathematics
and Mathematical Statistics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Max Sommerfeld
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Vladimir G. Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics (WIAS)
Mohrenstrasse 39
10117 Berlin
GERMANY

Gian Thanei
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Alexandre B. Tsybakov
ENSAE - CREST
Timbre J 340
3, Avenue Pierre Larousse
92240 Malakoff Cedex
FRANCE

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

Dr. Vincent Vu
Department of Statistics
The Ohio State University
325 Cakins Hall
1958 Neil Avenue
Columbus, OH 43210-1174
UNITED STATES

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

Yuting Wei
Department of Statistics
University of California, Berkeley
451 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Prof. Dr. Grace Yi
Faculty of Mathematics
Department of Statistics and
Actuarial Sciences
University of Waterloo
200 University Avenue W
Waterloo ON N2L 3G1
CANADA

Prof. Dr. Ming Yuan
Department of Statistics
University of Wisconsin
Medical Science Center
1300 University Avenue
Madison, WI 53706
UNITED STATES

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

Dr. Yoav Zemel
Section de Mathematiques
EPFL / MA-B1-493
Station 8
1015 Lausanne
SWITZERLAND

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