Mathematisches Forschungsinstitut Oberwolfach

Report No. 16/2016

DOI: 10.4171/OWR/2016/16

Computationally and Efficient Inference for Complex Large-scale Data

Organised by
Gilles Blanchard, Potsdam
Nicolai Meinshausen, Zürich
Richard Samworth, Cambridge
Ming Yuan, Madison

6 March – 12 March 2016

Abstract. The aim of the highly successful workshop *Computationally and statistically efficient inference for large-scale and heterogeneous data* was to foster dissemination and collaboration between researchers in the area of high-dimensional and large-scale data analysis. The field has grown tremendously over the last decade. Faced with ever larger data sets, many algorithms have emerged in computer science, machine learning and statistics that allow computationally efficient manipulation and model fitting on large datasets. Yet the mathematical and statistical properties of these algorithms are only just beginning to be understood. Advancing the field is important to avoid many misleading scientific discoveries based on pure data manipulation without the accompanying mathematical insights. The talks and discussions at the workshop covered the latest advances from optimization to statistical error control for large-scale data analysis.

Mathematics Subject Classification (2010): 62Gxx, 62Hxx, 62Jxx, 62C20, 68Q25, 68Q87, 68W15.

Introduction by the Organisers

The workshop *Computationally and statistically efficient inference for large-scale and heterogeneous data*, organised by Gilles Blanchard (Potsdam), Nicolai Meinshausen (ETH Zurich), Richard Samworth (Cambridge) and Ming Yuan (Madison) was well attended with 52 participants from a broad geographical background.

The background to the workshop is the data-driven revolution that most scientific fields are experiencing at the moment. Data are collected at an unprecented rate in most natural sciences. There was and still is early enthusiasm that the

flood of data will lead to a stream of new and interesting discoveries. However, one also expects (and observes in practice) many flawed results emerging from analyses that do not deal carefully enough with the statistical complexities inherent in the data. At the same time, many people are overwhelmed by the sheer amount of data and need procedures that combine computational efficiency with sound statistical inference.

Three broad themes and challenges associated with large-scale analysis of complex datasets were discussed at the workshop.

(1) **Tradeoffs and synergies between computational constraints, statistical efficiency, and optimization efficiency**. Datasets with millions or more of observations and variables imply that the computational efficiency of a statistical estimator is very important and many traditional approaches to inference are ruled out due to their inefficiency. An interesting question that is beginning to emerge is then whether one can characterize the statistically most efficient procedures under constraints on computational complexity. Another aspect of this theme is that one is naturally led to choose objective functions that leads to good statistical properties but can also be optimized reasonably fast. In such a context, the precise nature of the optimization can determine the number of samples that can be used in a specific analysis and efficient optimization can thus lead to more accurate estimates. Talks covering this topic included those of F. Bach, M. Drton, C. Heinze, C. Scott, G. Thanei, and T. Zhang.
(2) **Statistical error control for large-scale data**. Statistical inference is challenging in high-dimensional settings, for example if we are after causal effects of if the number of variables exceeds the number of variables. Regarding the latter problem: while a large body of results now exists on point estimators, it is only recently that the possibility of inference and confidence statements in these problems has emerged. In the context of linear models, most work has focused on confidence statements for regression parameters for high-dimensional linear models. Talks covering this topic included those of K. Balasubramanian, R. Foygel-Barber, T. Cai, J. Lei, A. Munk, J. Peters, and R. Tibshirani.
(3) **Statistics of complex structures for high-dimensional data**. The objects considered in traditional statistics are typically linear structures, classical parametric or nonparametric regression, density estimation in low-dimensional spaces. But with the explosion of the amount of available data comes a flurry of new scientific questions concerning statistical estimation and inference on increasingly complex structures (such as large-dimensional matrices, manifolds, trees, graphs or networks). While the fundamental statistical questions such as identifiability, consistency, limit behavior and convergence rates remain the same, the interaction with other areas of mathematics is developing at a fast rate and raises many new mathematical challenges. Talks covering this area included

those of R. Castro, C. Giraud, J. Lafferty, P-L. Loh, P. Rigollet, A. Rohde, D. Rothenhäusler, R. Willett, and Y. Yu.

These topics are of course strongly related. The second topic can for example be seen as a special case of the first topic: inference for high-dimensional linear models (part of the third theme) is a good example for the tradeoffs between statistical and computational efficiencies. For high-dimensional linear models, one typically uses convex penalties, as non-convex penalties such as constraints on the quasi-ℓ_0-norm of the regression vector are computationally infeasible for datasets with many variables. There is a statistical price to pay for the convexity of the objective function. While convex penalized estimation in high-dimensional problems allows for fast computation, it also incurs a bias in all estimates. Being able to quantify the bias and thus construct confidence-intervals for high-dimensional challenges has been an active focus of the workshop and will have many applications in scientific applications. Similarly, the second and the third topic are closely related, in particular concerning the fundational aspects of statistical error control for estimation of complex structures of various nature.

The workshop brought together a range of experts on the timely topic of how reliable statistical estimation and inference can be achieved in a computationally efficient manner, and how statistical methodology can be developed to handle complex structures arising in large-scale/large-dimensional data. We believe the talks and discussions held at the workshop will help to shape the field in the coming years.

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

Workshop: Computationally and Efficient Inference for Complex Large-scale Data

Table of Contents

Abstracts

"Harder, Better, Faster, Stronger" Convergence Rates for Least-Squares Regression

Francis Bach

(joint work with Aymeric Dieuleveut, Nicolas Flammarion)

Many supervised machine learning problems are naturally cast as the minimization of a smooth function defined on a Euclidean space. This includes least-squares regression, logistic regression (see, e.g., [1]) or generalized linear models [2]. While small problems with few or low-dimensional input features may be solved precisely by many potential optimization algorithms (e.g., Newton method), large-scale problems with many high-dimensional features are typically solved with simple gradient-based iterative techniques whose per-iteration cost is small.

In this paper, we consider a quadratic objective function f whose gradients are only accessible through a stochastic oracle that returns the gradient at any given point plus a zero-mean finite variance random error. In this stochastic approximation framework [3], it is known that two quantities dictate the behavior of various algorithms, namely the covariance matrix V of the noise in the gradients, and the deviation $\theta_0 - \theta_*$ between the initial point of the algorithm θ_0 and any of the global minimizer θ_* of f. This leads to a "bias/variance" decomposition [4, 5] of the performance of most algorithms as the sum of two terms: (a) the bias term characterizes how fast initial conditions are forgotten and thus is increasing in a well-chosen norm of $\theta_0 - \theta_*$; while (b) the variance term characterizes the effect of the noise in the gradients, independently of the starting point, and with a term that is increasing in the covariance of the noise.

For quadratic functions with (a) a noise covariance matrix V which is proportional (with constant σ^2) to the Hessian of f (a situation which corresponds to least-squares regression) and (b) an initial point characterized by the norm $\|\theta_0 - \theta_*\|^2$, the optimal bias and variance terms are known *separately*. On the one hand, the optimal bias term after n iterations is proportional to $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$, where L is the largest eigenvalue of the Hessian of f. This rate is achieved by accelerated gradient descent [6, 7], and is known to be optimal if the number of iterations n is less than the dimension d of the underlying predictors, but the algorithm is not robust to random or deterministic noise in the gradients [8, 9]. On the other hand, the optimal variance term is proportional to $\frac{\sigma^2 d}{n}$ [10]; it is known to be achieved by averaged gradient descent [4], which for the bias term only achieves $\frac{L\|\theta_0-\theta_*\|^2}{n}$ instead of $\frac{L\|\theta_0-\theta_*\|^2}{n^2}$.

Our first contribution in this paper is to present a novel algorithm which attains optimal rates for *both the variance and the bias terms*. This algorithm is averaged accelerated gradient descent; beyond obtaining jointly optimal rates, our result shows that averaging is beneficial for accelerated techniques and provides a provable robustness to noise.

While optimal when measuring performance in terms of the dimension d and the initial distance to optimum $\|\theta_0 - \theta_*\|^2$, these rates are not adapted in many situations where either d is larger than the number of iterations n (i.e., the number of observations for regular stochastic gradient descent) or $L\|\theta_0 - \theta_*\|^2$ is much larger than n^2. Our second contribution is to provide an analysis of a new algorithm (based on some additional regularization) that can adapt our bounds to finer assumptions on $\theta_0 - \theta_*$ and the Hessian of the problem, leading in particular to dimension-free quantities that can thus be extended to the Hilbert space setting [11] (in particular for non-parametric estimation). For more details, see [12].

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition, 2009.

[2] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, second edition, 1989.

[3] H. Robbins and S. Monro. A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407, 1951.

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, December 2013.

[5] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[6] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[7] Y. Nesterov. *Introductory Lectures on Convex Optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[8] A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19 (3):1171–1183, 2008.

[9] O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1-2, Ser. A):37–75, 2014.

[10] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

[11] A. Dieuleveut and F. Bach. Non-parametric Stochastic Approximation with Large Step sizes. Technical report, ArXiv, 2014.

[12] A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. Technical Report 1602.05419, arXiv, 2016.

Optimal and Adaptive Goodness-of-fit Testing via Kernel Methods

Krishna Balasubramanian

(joint work with Ming Yuan)

Whenever a statistical model is posed, it is crucial to check its validity. More specifically, let $X_1, \ldots, X_n$ be independent observations sampled from an unknown distribution P on a measurable space on $(\mathcal{X}, \mathcal{B})$. We wish to test a null hypothesis $H_0 : P = P_0$, where P_0 is some fixed distribution on $(\mathcal{X}, \mathcal{B})$. This problem, often referred to as testing for goodness-of-fit, has a long and illustrious history in statistics and is often associated with household names such as *Kolmogrov-Smirnov*

tests, Pearson's Chi-square test or *Neyman's smooth test*. A plethora of other techniques have also been proposed over the years, both in the parametric setting and the non-parametric setting. Most of the existing techniques are developed with the domain $\mathcal{X} = \mathbb{R}$ or $[0, 1]$ in mind and work the best in these cases. Modern applications, however, oftentimes involve domains different from these traditional ones.

A particularly attractive approach to goodness-of-fit testing problems in general domains is through the reproducing kernel Hilbert space (RKHS) embedding of distributions, which has attracted a lot of attention in recent years. Like other kernel methods, RKHS embedding based tests present a general and unifying framework for goodness-of-fit testing problems in arbitrary domains by using appropriate kernels defined on those domains. More specifically, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel that is symmetric, positive (semi-)definite and square integrable. The RKHS embedding of a probability measure P on $(\mathcal{X}, \mathcal{B})$, with respect to K, is given by

$$\mu_P(\cdot) := \int_{\mathcal{X}} K(x, \cdot) P(dx).$$

The Moore-Aronszajn Theorem indicates that there is a RKHS, denoted by $(\mathcal{H}(K), \langle \cdot, \cdot \rangle_K)$, uniquely identified with the kernel. It is clear that $\mu_P \in \mathcal{H}(K)$, and hence the notion of RKHS embedding. Based on this, the so called *maximum mean discrepancy* (MMD) between two probability measures P and Q is defined as

$$\gamma_K(P, Q) = \|\mu_P - \mu_Q\|_K.$$

The goodness-of-fit test can be carried out conveniently through RKHS embeddings of P and P_0 by first constructing an estimate of $\gamma_K(P, P_0)$.

Despite its popularity, little is known about the performance of the aforementioned RKHS embedding based goodness-of-fit test. Our goal is to fill in this void. In this work, we investigate the power of the above discussed testing strategy under a general composite alternative. We are particularly interested in the detection boundary, namely how close P and P_0 can be (with respect to χ^2 distance), under the alternative, so that a test can still consistently distinguish between the null hypothesis and the alternative. Our first result suggests that the detection boundary for $\hat{\gamma}_n(P, P_0)$ is of the order $n^{-1/2}$. It is of interests to compare this rate with those typically achieved in a parametric setting where it is known that, in general, consistent tests are available whenever the detection boundary is of the order n^{-1}. A natural question is to what extent such a gap can be attributed to the fundamental difference between parametric and nonparametric testing problems. It turns out that much of it is actually due to the suboptimality of MMD, and the rates attained by a test using MMD can be significantly improved through a slight modification of the MMD. For concreteness, we further assume that the eigenvalues of K with respect to $L_2(P_0)$ decays polynomially in that $\lambda_k \asymp k^{-2s}$. We show that the critical radius for testing H_0 against H_1 for any $\theta \geq 0$ is $n^{-\frac{4s}{4s+\theta+1}}$. The rate of detection can be achieved, in particular, by a moderated version of the MMD based approach. A practical challenge to the approach, however, is its reliance

on the knowledge of θ. Unlike s which is determined by K and P_0 and therefore known apriori, θ depends on u and is not known in advance. This naturally brings about the issue of adaptation – is there a single test that can adaptively attain the optimal detection boundary without the knowledge of θ. We show that the answer is affirmative although a small price in the form of $\log\log n$ is required to achieve such adaptation.

High dimensional inference and sign errors with knockoffs

Rina Foygel Barber
(joint work with Emmanuel J. Candès)

We develop a framework for testing for associations in a possibly high-dimensional linear model where the number of features p may far exceed the sample size n. We consider the model $y = X\beta + \epsilon$, where $y \in \mathbb{R}^n$ is the real-valued response, $X \in \mathbb{R}^{n\times p}$ is the design matrix consisting of p features (the p columns of X), and $\epsilon \sim N(0, \sigma^2\mathbf{I}_n)$ is the noise, assumed to be i.i.d. Gaussian. Our goal is to select a model, that is, a set of features $\widehat{S} \subset \{1, \dots, p\}$, while controlling the proportion of false discoveries in the model: defining $\mathcal{H}_0 = \{j : \beta_j = 0\}$, the set of "null" features whose true coefficient is zero, we would like to bound the false discovery rate (FDR), i.e. the expected false discovery proportion (FDP), defined as

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] = \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{H}_0|}{|\widehat{S}| \vee 1}\right].$$

We may also consider sign error rates, known as the directional FDR [3], where we also count as an error, any feature whose true effect β_j is nonzero but was selected with the wrong sign—these errors are known as "Type S" [4]. Formally, we partition our selected set of features as $\widehat{S} = \widehat{S}_+ \cup \widehat{S}_-$, where $\widehat{S}_+$ is the set of features in the selected model that we estimate to have a positive effect, and $\widehat{S}_-$ is the same for negative effects. We then have the directional FDR:

$$\mathrm{FDR}_{\mathrm{dir}} = \mathbb{E}[\mathrm{FDP}_{\mathrm{dir}}] = \mathbb{E}\left[\frac{|\{j \in \widehat{S}_+, \beta_j \le 0\}| + |\{j \in \widehat{S}_-, \beta_j \ge 0\}|}{|\widehat{S}| \vee 1}\right].$$

The directional FDR is always at least as large as the FDR, since we now are penalized for Type S errors in addition to Type I errors.

Our method, the knockoff filter, controls the FDR and the directional FDR, and can be applied in either a low dimensional ($n \ge p$) or high dimensional ($n < p$) setting, although for high dimensions we work with inference within a reduced model, described below. First, we discuss the low dimensional setting. Consider a model selection method such as the Lasso [5], which solves

$$\widehat{\beta}(\lambda) = \arg\min_{b\in\mathbb{R}^p} \left\{\frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1\right\},$$

where $\lambda > 0$ is a penalty parameter with large values ensuring a sparse fitted model. As λ moves from $+\infty$ to 0, this produces a sequence of features entering

the path at different times. We could also consider a forward stepwise method, or other methods producing a path of nested models. At any point in the path, a feature X_j which is a null (i.e. $\beta_j = 0$) may be selected for one of two reasons: either (1) it is correlated with the noise ϵ, or (2) it is correlated with some true signal X_k which has been missed by the model path (or, relatedly, the coefficient β_k has been underestimated, perhaps due to shrinkage). In order to estimate the number of false positives along the model path, then, we need to account for both of these sources of error.

To do so, our method creates a knockoff copy—a fake variable serving as a control—for each of the p features; the knockoff copy for X_j is denoted by $\widetilde{X}_j$. By requiring the knockoffs to satisfy certain correlation conditions mimicking the original features' correlation structure, these knockoffs then act as a "control group" for any path of nested models produced by the chosen model selection method, and are equally likely as their corresponding null features to be falsely selected in the model path, for either of the two reasons mentioned above.

We then apply the model selection method to the augmented data set consisting of the response y and the $2p$ features $X_1, \ldots, X_p, \widetilde{X}_1, \ldots, \widetilde{X}_p$. At each point in the model path, we estimate the current number of false positives by examining the number of knockoff features that have been selected so far. The knockoff filter then stops at the last time when the resulting estimated false discovery proportion is still below some chosen threshold. Specifically, at each time λ in the model path (moving from the empty model at $\lambda = +\infty$ to the full model at $\lambda = 0$), define

$$S(\lambda) = \{j \ : \ \text{feature } X_j \text{ entered the path by time } \lambda, \text{ and before its knockoff } \widetilde{X}_j\}$$

and the corresponding "mirror image"

$$\widetilde{S}(\lambda) = \{j \ : \ \text{knockoff } \widetilde{X}_j \text{ entered the path by time } \lambda, \text{ and before its original } X_j\}.$$

Since the knockoff copy $\widetilde{X}_j$ and its corresponding original null feature X_j are equally likely to be (falsely) selected, and equally likely to enter the path in either order, the number of selected null knockoffs, $|\widetilde{S}(\lambda) \cap \mathcal{H}_0|$, is an estimate of the number of selected null original features, $|S(\lambda) \cap \mathcal{H}_0|$. Therefore we can estimate the FDP at the current point in the path as $\mathrm{FDP}(S(\lambda)) \approx |\widetilde{S}(\lambda)|/|S(\lambda)|$. We then choose the smallest λ (i.e. the largest model) such that this estimated FDP is bounded by some desired level q, e.g. $q = 0.1$; the final output is the selected model $S(\lambda)$. Our theoretical results for low dimensions prove that the modified directed FDR is bounded,

$$\mathrm{mFDR}_{\mathrm{dir}} = \mathbb{E}\left[\frac{|\{j \in \widehat{S}_+, \beta_j \le 0\}| + |\{j \in \widehat{S}_-, \beta_j \ge 0\}|}{|\widehat{S}| + q^{-1}}\right] \le q.$$

The knockoff filter can be applied also in high dimensions ($n < p$): the observations are split into two groups, where the first group is used to screen for a set $S \subset \{1, \ldots, p\}$ of potentially relevant variables, whereas the second is used for inference over this reduced set of variables; we also develop strategies for leveraging information from the first part of the data at the inference step for greater

accuracy. After the screened model is chosen, the inferential step is carried out with the low-dimensional knockoff filter.

The simplest form of this approach is to use data splitting, where we partition our n observations into n_0 observations $(X^{(0)}, y^{(0)})$ used for screening, and $n_1 = n - n_0$ observations $(X^{(1)}, y^{(1)})$ used for inference. However, at the inference step, restricting ourselves to the reduced data $(X^{(1)}, y^{(1)})$ lowers the power of our method to detect true signals; ideally we would want to reuse the first part of the data, $(X^{(0)}, y^{(0)})$, again for the inference step even though it was already used at the screening step. To do so, we first create knockoff features for $X_S^{(1)}$, denoted as $\widetilde{X}_S^{(1)}$. We then apply the knockoff filter using the data

$$X_S = \begin{pmatrix} X_S^{(0)} \\ X_S^{(1)} \end{pmatrix}, \ \widetilde{X}_S = \begin{pmatrix} X_S^{(0)} \\ \widetilde{X}_S^{(1)} \end{pmatrix}, \ y = \begin{pmatrix} y^{(0)} \\ y^{(1)} \end{pmatrix}.$$

Note that the knockoff features, i.e. the columns of $\widetilde{X}_S$, are exactly equal to their corresponding original features on the first n_0 observations; they only differ on the last n_1 observations. This means that, when we compare for instance $X_j^\top y$ against $\widetilde{X}_j^\top y$ for some screened feature $j \in S$, the difference depends on y only through $y^{(1)}$, the part of the response which has not been observed yet as it was not used for the screening step. This "data recycling" trick allows the inference properties of the knockoff filter, i.e. the results on FDR and directional FDR control, to be maintained even though we are reusing the first part of the data.

In high dimensions, since the screening step may potentially miss some true signals, the directional FDR control results apply to the partial regression coefficients in the reduced model consisting of only those features selected by the screening step; that is, if $S \subseteq \{1, \ldots, p\}$ is the screened set of features, then we are performing inference on $\beta^S \in \mathbb{R}^S$, the partial regression coefficients when the response y is regressed on the set of features $\{X_j : j \in S\}$.

Finally, we demonstrate the performance of our approach through numerical studies showing more power than existing alternatives, and we also apply our method to a genome-wide association study to find locations on the genome that are possibly associated with continuous phenotypes (LDL and HDL levels).

Our work on this method appears in [1] for low dimensions proving FDR control; subsequent work in [2] proves directional FDR control in low dimensions and extends to the high dimensional setting.

References

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

[2] R. F. Barber and E. J. Candès, *A knockoff filter for high-dimensional selective inference.* Preprint, arXiv:1602.03574 (2016).

[3] Y. Benjamini and D. Yekutieli, *False discovery rate-adjusted multiple confidence intervals for selected parameters.* Journal of the American Statistical Association **100** (2005), 71–81.

[4] A. Gelman and F. Tuerlinckx, *Type S error rates for classical and Bayesian single and multiple comparison procedures.* Computational Statistics **15** (2000), 373–390.
[5] R. Tibshirani, *Regression shrinkage and selection via the Lasso.* Journal of the Royal Statistical Society, Series B **58** (1996), 267–288.

Distribution-Free Detection of Structured Anomalies: Permutation and Rank-Based Scans

Rui M. Castro
(joint work with Ery Arias-Castro, Ervin Tánczos and Meng Wang)

The scan statistic is by far the most popular method for anomaly detection, and is used often for syndromic surveillance, signal and image processing, and target detection based on sensor networks. Tests based on the scan statistics are easily calibrated with the knowledge the null distribution, corresponding to the absence of abnormal behavior. When this distribution is unknown it is less clear how to proceed. We investigate two possible approaches: (i) calibration by permutation and; (ii) a rank-scan test, which is distribution-free and less sensitive to outliers. Furthermore, knowledge of the sample size suffices to calibrate the rank-scan test, so computationally it has the same cost as the oracle scan test. In both cases, we quantify the performance loss with respect to an oracle scan test (with full knowledge of the null distribution) and show there is only a small loss of power in the context of a natural exponential family. This includes the classical normal location model, popular in signal processing, and the Poisson model, popular in syndromic surveillance. An extended version of this report can be found in [2].

Specifically, we observe a realization $\mathbf{x}$ of a set independent random variables $\mathbf{X} \equiv (X_i : i \in [N])$, where N is the dataset size and $[N] \equiv \{1, \ldots, N\}$. We take a binary hypothesis testing point of view. Under the null hypothesis these random variables are identically distributed with some unknown null distribution F_0. Under the alternative, a subset of these variables has a different distribution. Let $\mathbb{S} \subset 2^{[N]}$ denote a class of possibly anomalous subsets. Under the alternative hypothesis there is a subset $\mathcal{S} \in \mathbb{S}$ such that, for each $i \in \mathcal{S}$, $X_i \sim F_i$ for some distribution $F_i \neq F_0$, while $(X_i : i \in [N] \setminus \mathcal{S})$ still have distribution F_0. In a number of important applications the variables are real-valued and the anomalous observations take larger-than-usual values, formalized by assuming that each F_i stochastically dominates F_0. Without loss of generality and to simplify the presentation below we assume F_0 has zero mean and variance one for the results in Theorem 1.1 and Corollary 2.2.

When the distributions are known a test based on the generalized likelihood ratio is a very natural approach. An asymptotic equivalent alternative is to use the scan statistic,

$$\mathrm{SCAN}(\mathbf{x}) = \max_{\mathcal{S} \in \tilde{\mathbb{S}}} \frac{1}{\sqrt{|\mathcal{S}|}} \left(\sum_{i \in \mathcal{S}} x_i - \frac{1}{N} \sum_{i \in [N]} x_i \right) ,$$

where $\tilde{\mathbb{S}}$ is surrogate for $\mathbb{S}$ (typically an approximating net). The use of a surrogate for $\mathbb{S}$ is both motivated by computational and analytical considerations. A test based on this statistic rejects the null hypothesis if $\mathrm{SCAN}(\mathbf{x})$ is larger than some threshold. If the null distribution is known the corresponding test can be calibrated by Monte-Carlo simulation. If the distribution is not known two possible alternatives come to mind, namely to calibrate the test by permutation, or replace the observations by the corresponding ranks and use a scan statistic over the ranks instead.

1. Calibration by Permutation

The p-value of the scan test calibrated by permutation is defined as

$$\mathfrak{P}(\mathbf{x}) = \frac{1}{N!}\Big|\big\{\pi \in [N]! : \mathrm{SCAN}(\mathbf{x}_\pi) \geq \mathrm{SCAN}(\mathbf{x})\big\}\Big| \,, \tag{1}$$

where $[N]!$ denotes the set of all permutations of $[N]$. A test of level $\alpha \in (0,1)$ rejects the null hypothesis if $\mathfrak{P}(\mathbf{x}) \leq \alpha$. The statistical properties of such a test depend obviously on the class $\mathbb{S}$ and the unknown underlying distributions. For concreteness we consider in this report the class of intervals

$$\mathbb{S} = \big\{\{a, \ldots, b\} : 1 \leq a \leq b \leq N\big\} \ .$$

Theorem 1.1. *The permutation scan test defined above has level at most α. Assume further that F_i belongs to a one-parameter exponential family in natural form, so that the Radon-Nikodym derivative of F_i with respect to F_0 is given by $f_i(x) \propto \exp(\theta_i x)$, where $\theta_i \in [0, \theta^*)$. Under the alternative, and provided $|\mathcal{S}| = o(N)$ and $|\mathcal{S}|/\log^3 N \to \infty$ the permutation scan test has power converging to one*[1] *as $N \to \infty$ when*

$$\min_{i \in \mathcal{S}} \theta_i \geq \tau \sqrt{\frac{2 \log N}{|S|}}$$

with $\tau > 1$.

Furthermore, the (oracle) scan test calibrated with the full knowledge of the distribution has precisely the same characterization, while when $\tau < 1$ there is no powerful test even with the full knowledge of the distributions [1]. The conclusion is that calibration by permutation has the same asymptotic power as the optimal test (to first-order accuracy). This type of result and our analysis methodology extends naturally to other classes $\mathbb{S}$ as well.

2. The Rank-Scan Test

Let R_i denote the rank (in increasing order) of X_i among $\mathbf{X}$. Ties are broken randomly to ensure the distribution of $\mathbf{R} = \{R_i,\ i \in [N]\}$ under the null is the permutation distribution. The rank-scan test statistic is simply $\mathrm{SCAN}(\mathbf{r})$, leading

[1] A rather technical assumption is also required, namely either F_0 has compact support or $\max_i \theta_i \leq \bar{\theta} < \theta^*$ for some fixed $\bar{\theta} > 0$. It is possible to drop this assumption by censoring the observations before computing the scan statistic.

to a distribution-free test. Recalling equation (1) the p-value of the rank-scan test is given by $\mathfrak{P}(\mathbf{r})$ and we reject the null hypothesis at level α if $\mathfrak{P}(\mathbf{r}) \leq \alpha$. Define

$$p_i = \mathbb{P}(Y > X) + \frac{1}{2}\mathbb{P}(Y = X) \ ,$$

where X and Y are independent random variables with distributions F_0 and F_i respectively.

Theorem 2.1. *The rank-scan test defined above has level at most α. Furthermore it has power converging to one as $N \to \infty$ if:*

(1) $|\mathcal{S}|/\log N \to \infty$, $|\mathcal{S}| = o(N)$ and

$$\min_{i \in \mathcal{S}} p_i \geq \frac{1}{2} + \tau \sqrt{\frac{2 \log N}{|S|}} \text{ with } \tau > \frac{1}{2\sqrt{3}}$$

(2) $|\mathcal{S}| = c \log N$ for some $c > 0$ and

$$\min_{i \in \mathcal{S}} p_i \geq 1 - \frac{1}{2} \exp\left(-\frac{c+1}{c}\right)$$

(3) $2 < |\mathcal{S}| = o(\log N)$ and

$$\min_{i \in \mathcal{S}} p_i = 1 - o\left((N \log N)^{-2/|S|}\right) \ ,$$

This result holds for arbitrary distributions. For the distributions in the one-parameter exponential family as in Theorem 1.1 we have the following result.

Corollary 2.2. *In the setting of Theorem 1.1 the rank-scan test is asymptotically powerful provided*

$$\min_{i \in \mathcal{S}} \theta_i \geq \tau \sqrt{\frac{2 \log N}{|S|}}$$

with $\tau > \frac{1}{\sqrt{3}\Upsilon_0}$ and $\Upsilon_0 = \mathbb{E}[\max(X, Y)]$, where X, Y are independent random variables with distribution F_0.

The parameter Υ_0 characterizes the loss of power of the rank-scan test as a function of the null distribution. If F_0 is uniform then $1/(\sqrt{3}\Upsilon_0) = 1$ so there is asymptotically no loss of power with respect to the oracle scan test. For the emblematic case of the normal location model where F_0 is the standard normal distribution $1/(\sqrt{3}\Upsilon_0) = \sqrt{\pi/3} \approx 1.023$, meaning the rank-scan test has almost no loss of power in comparison with the oracle scan test. In addition, numerical experiments with synthetic and real data further confirm these theoretical findings and demonstrate the very good performance of the rank-scan test in finite sample scenarios [2].

References

[1] Arias-Castro E., Candès E. J. and Durand A., *Detection of an anomalous cluster in a network*, The Annals of Statistics **39(1)** (2011), 278–304.

[2] Arias-Castro E., Castro R. M., Tánczos E. and Wang M., *Distribution-Free Detection of Structured Anomalies: Permutation and Rank-Based Scans*, arXiv preprint arXiv:1508.03002.

Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity

Tony Cai

(joint work with Zijian Guo)

Consider the high-dimensional linear regression model

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathrm{I}), \tag{1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. Several penalized/constrained ℓ_1 minimization methods, including the Lasso, Dantzig Selector, square-root Lasso, and scaled Lasso have been proposed and studied. Under regularity conditions on the design matrix X, these methods with a suitable choice of the tuning parameter have been shown to achieve the optimal rate of convergence $k\frac{\log p}{n}$ under the squared error loss over the set of k-sparse regression coefficient vectors with $k \leq c\frac{n}{\log p}$ where $c > 0$ is a constant. See, for example, [5, 1]. A key feature of the estimation problem is that the optimal rate can be achieved adaptively with respect to the sparsity parameter k.

Confidence sets play a fundamental role in statistical inference and confidence intervals for high-dimensional linear regression have been actively studied recently with a focus on inference for individual coordinates. Zhang and Zhang [6] was the first to use de-biasing for constructing a confidence interval for β_i. [2, 3, 4] also used de-biasing for the construction of confidence intervals and [4] established asymptotic efficiency for the proposed estimator. All the aforementioned papers [6, 2, 3, 4] have focused on the ultra-sparse case where the sparsity $k \ll \frac{\sqrt{n}}{\log p}$ is assumed. Under such a sparsity condition, the expected length of the confidence intervals constructed in [6, 3, 4] is at the parametric rate $\frac{1}{\sqrt{n}}$ and the procedures do not depend on the specific value of k.

Compared to point estimation where the sparsity condition $k \ll \frac{n}{\log p}$ is sufficient for estimation consistency, the condition $k \ll \frac{\sqrt{n}}{\log p}$ for valid confidence intervals is much stronger. There are several natural questions: What happens in the region where $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$? Is it still possible to construct a valid confidence interval for β_i? Can one construct an adaptive honest confidence interval not depending on k? Our goal is to address these and other related questions. Specifically, we consider confidence intervals for a linear functional $\mathrm{T}(\beta) = \xi^\intercal \beta$, where the loading vector $\xi \in \mathbb{R}^p$ is given and $\frac{\max_{i\in \mathrm{supp}(\xi)} |\xi_i|}{\min_{i\in \mathrm{supp}(\xi)} |\xi_i|} \leq \bar{c}$ with $\bar{c} \geq 1$ being a constant. Based on the sparsity of ξ, we focus on two specific regimes: the sparse loading regime

where $\|\xi\|_0 \leq Ck$, with $C > 0$ being a constant; the dense loading regime where $\|\xi\|_0 \gg k^2$. For confidence intervals, $\mathrm{T}(\beta) = \beta_i$ is a prototypical case for the general functional $\mathrm{T}(\beta) = \xi^\intercal \beta$ with a sparse loading ξ, and $\mathrm{T}(\beta) = \sum_{i=1}^p \beta_i$ is a representative case for $\mathrm{T}(\beta) = \xi^\intercal \beta$ with a dense loading ξ.

We first focus on the two specific functionals $\mathrm{T}(\beta) = \beta_i$ and $\mathrm{T}(\beta) = \sum_{i=1}^p \beta_i$. We establish the convergence rate of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter k is given. It is shown that in this case the minimax expected length is of order $\frac{1}{\sqrt{n}} + k\frac{\log p}{n}$ for confidence intervals of β_i. An honest confidence interval, which depends on the sparsity k, is constructed and is shown to be minimax rate optimal. To the best of our knowledge, this is the first construction of confidence intervals in the moderate-sparse region $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$. If the sparsity k falls into the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, the constructed confidence interval is similar to the ones constructed in [6, 3, 4]. On the other hand, the convergence rate of the minimax expected length of confidence intervals for $\sum_{i=1}^p \beta_i$ in the oracle setting is shown to be $k\sqrt{\frac{\log p}{n}}$. A rate-optimal confidence interval that also depends on k is constructed. It should be noted that this confidence interval is not based on the de-biased estimator.

One drawback of the confidence intervals mentioned above is that they require knowing the sparsity k. Such knowledge of sparsity is usually unavailable in applications. A natural question is: Without knowing the sparsity k, is it possible to construct a confidence interval as good as when the sparsity k is known? This is a question about adaptive inference, which has been a major goal in nonparametric and high-dimensional statistics. Ideally, an adaptive confidence interval should have its length automatically adjusted to the true sparsity of the unknown regression vector, while maintaining a prespecified coverage probability. We show that, in marked contrast to point estimation, such a goal is in general not attainable for confidence intervals. In the case of confidence intervals for β_i, it is impossible to adapt between different sparsity levels, except when the sparsity k is restricted to the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, over which the confidence intervals have the optimal length of the parametric rate $\frac{1}{\sqrt{n}}$, which does not depend on k. In the case of confidence intervals for $\sum_{i=1}^p \beta_i$, it is shown that adaptation to the sparsity is not possible at all, even in the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$.

Minimax theory is often criticized as being too conservative as it focuses on the worst case performance. For confidence intervals for high dimensional linear regression, we establish strong non-adaptivity results which demonstrate that the lack of adaptivity is not due to the conservativeness of the minimax framework. It shows that for any confidence interval with guaranteed coverage probability over the set of k sparse vectors, its expected length at any given point in a large subset of the parameter space must be at least of the same order as the minimax expected length. So the confidence interval must be long at a large subset of points in the parameter space, not just at a small number of "unlucky" points. This leads directly to the impossibility of adaptation over different sparsity levels.

Fundamentally, the lack of adaptivity is caused by the difficulty in accurately learning the bias of any estimator for high-dimensional linear regression.

We now turn to confidence intervals for general linear functionals. For a linear functional $\xi^\intercal\beta$ in the sparse loading regime, the rate of the minimax expected length is $\|\xi\|_2\left(\frac{1}{\sqrt{n}}+k\frac{\log p}{n}\right)$, where $\|\xi\|_2$ is the vector ℓ_2 norm of ξ. For a linear functional $\xi^\intercal\beta$ in the dense loading regime, the rate of the minimax expected length is $\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}$, where $\|\xi\|_\infty$ is the vector ℓ_∞ norm of ξ. Regarding adaptivity, the phenomena observed in confidence intervals for the two special linear functionals $\mathrm{T}(\beta)=\beta_i$ and $\mathrm{T}(\beta)=\sum_{i=1}^p\beta_i$ extend to the general linear functionals. The case of confidence intervals for $\mathrm{T}(\beta)=\sum_{i=1}^p\xi_i\beta_i$ with a sparse loading ξ is similar to that of confidence intervals for β_i in the sense that rate-optimal adaptation is impossible except when the sparsity k is restricted to the ultra-sparse region $k\lesssim\frac{\sqrt{n}}{\log p}$. On the other hand, the case for a dense loading ξ is similar to that of confidence intervals for $\sum_{i=1}^p\beta_i$: adaptation to the sparsity k is not possible at all, even in the ultra-sparse region $k\lesssim\frac{\sqrt{n}}{\log p}$.

In addition to the more typical setting in practice where the covariance matrix Σ of random design and the noise level σ of the linear model are unknown, we also consider the case with the prior knowledge of $\Sigma=\mathrm{I}$ and $\sigma=\sigma_0$. It turns out that this case is strikingly different. The minimax rate for the expected length in the sparse loading regime is reduced from $\|\xi\|_2\left(\frac{1}{\sqrt{n}}+k\frac{\log p}{n}\right)$ to $\frac{\|\xi\|_2}{\sqrt{n}}$, and in particular it does not depend on the sparsity k. Furthermore, in marked contrast to the case of unknown Σ and σ, adaptation to sparsity is possible over the full range $k\lesssim\frac{n}{\log p}$. On the other hand, for linear functionals $\xi^\intercal\beta$ with a dense loading ξ, the minimax rates and impossibility for adaptive confidence intervals do not change even with the prior knowledge of $\Sigma=\mathrm{I}$ and $\sigma=\sigma_0$. However, the cost of adaptation is reduced with the prior knowledge.

References

[1] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[2] Adel Javanmard and Alessandro Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *Information Theory, IEEE Transactions on*, 60(10):6522–6554, 2014.

[3] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[4] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[5] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[6] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Estimation of High-Dimensional Graphical Models Using Regularized Score Matching

Mathias Drton
(joint work with Lina Lin, Ali Shojaie)

We discuss estimation of undirected conditional independence graphs for high-dimensional continuous random vectors. In such a graph, nodes correspond to random variables, and an edge is present if two variables are conditionally dependent given all other variables. For the estimation, we consider the application of the score matching approach, introduced and subsequently extended by Hyvärinen [1, 2]. The *regularized score matching* method we propose allows for computationally efficient treatment of possibly non-Gaussian exponential family models. Indeed, the score matching loss is a convex quadratic function for any exponential family of continuous distributions, which offers great flexibility in model specification while keeping computation as well as theoretical analysis tractable. One particular consequence of the quadratic nature of the loss is that score matching admits piecewise linear solution paths under ℓ_1 regularization.

In the well-explored Gaussian setting, regularized score matching yields symmetric estimates of a sparse precision matrix. As we demonstrate, the method is state-of-the-art in this case. The true potential of the method, however, lies in its application to non-Gaussian models. Particular instances we treat are models of distributions with Gaussian conditionals and Gaussian distribution truncated to the nonnegative orthant. In the latter case, we apply the extension from [2]. Under suitable irrepresentability conditions, we show that ℓ_1-regularized score matching is consistent for graph estimation in sparse high-dimensional settings. Through simulation experiments, we demonstrate good performance of regularized score matching also for non-Gaussian settings. In an application to RNAseq data, we show that exploratory analysis based on truncated Gaussian models can extract information in a way that is very complementary to what is obtained from existing methods.

References

[1] A. Hyvärinen, *Estimation of non-normalized statistical models by score matching*, J. Mach. Learn. Res. **6** (2005), 695–709.

[2] A. Hyvärinen, *Some extensions of score matching*, Comput. Statist. Data Anal. **51** (2007), 2499–2512.

Variable clustering with *G*-models

Christophe Giraud
(joint work with Florentina Bunea, Xi Luo, Martin Royer and Nicolas Verzelen)

While widely used in applied statistics, the problem of *variable clustering* has attracted little attention from a theoretical point of view. In the papers [1, 2], we propose some new probabilistic modelings and some new algorithms for variable clustering and we prove some optimality under various settings.

The purpose of variable clustering is to cluster the entries of a p-dimensional vector $X = (X_1, \dots, X_p)$ into groups $(X_a)_{a\in G_1}, \dots, (X_a)_{a\in G_K}$, such that within group variables are *similar*. The implicit notion of similarity underlying classical clustering algorithm like K-means or ℓ^2 Hierarchical clustering is that *within-group variables are more correlated than between group variables*. Instead, we propose a family of probabilistic models, called G-models, formalizing the principle that *within group variables behave similarly with respect to all the other variables*.

In G-models, the random vector X is assumed to have a mean 0 and a covariance matrix Σ. A simple and natural notion to encode the above principle, is that *switching two variables of a same group does not change the distribution of the vector* X. This notion of partial exchangeability has the nice property to lead to a non-ambiguous definition of the groups. Actually, there exists a unique minimal partition such that within group variables can be permuted without altering the distribution of X. Yet, this clustering criterion is not easily amenable to estimation from a n-sample of X. An idea is then to relax this modeling, by merely looking at the consequence of partial exchangeability on the covariance matrix Σ.

Let $G = \{G_1, \dots, G_K\}$ be a partition of $\{1, \dots, p\}$ such that within group variables can be permuted without altering the distribution of X. Let us write $k(a)$ for the index k of the group such that $X_a \in G_k$. Then, the covariance Σ fulfills that

- the off-diagonal entries Σ_{ab} depend only on the indices $k(a)$ and $k(b)$;
- the diagonal entries Σ_{aa} depend only on the indices $k(a)$.

Hence, the covariance matrix can be decomposed into

$$\Sigma = ACA^T + \Gamma, \tag{1}$$

where the $p \times K$ matrix $A_{ak} = 1_{\{a\in G_k\}}$ assigns the index of a variable X_a to a group G_k, the matrix C is symmetric, and Γ is a diagonal matrix with $\Gamma_{aa} = \gamma_k$ for all $a \in G_k$. A matrix Σ fulfilling such a decomposition with respect to a partition G is said to be G-block structured. Again, we have an identifiable target partition for clustering with respect to this property, since there exists a unique partition G^* according to which Σ fulfills the decomposition (1). We emphasize that the matrix C may *not* be positive semi-definite and that the within-group correlations may be smaller than the between group correlations, as illustrated in the following

example

$$\Sigma = \begin{bmatrix} 1 & -1/2 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & -1/2 & 1 \end{bmatrix}. \tag{2}$$

For simplicity, we assume from now on that X is a Gaussian random variable. All results can be readily extended to Gaussian copulas or sub-gaussian random variables.

The partition G^* can be retrieve from n i.i.d. observations of X, only if the dissimilarity between the groups is strong enough. In G-models, this dissimilarity is caught via the CORD metric

$$CORD(a,b) = \max_{c \neq a,b} |\Sigma_{ac} - \Sigma_{bc}|$$

which compares the covariance of a and b with respect to all the other variables. Bunea et al. [1] propose a computationally efficient algorithm retrieving G^* with high-probability when the block separation condition

$$MCORD(\Sigma) := \min_{a \overset{G^*}{\not\sim} b} CORD(a,b) \geq c\sqrt{\frac{\log(p)}{n}}$$

is met. The rate $\sqrt{\log(p)/n}$ in the above block separation condition is minimax optimal for perfect recovery, in the sense that there exists a constant c^* such that

$$\inf_{\hat{G}} \sup \left\{ \mathbf{P}_\Sigma(\hat{G} \neq G^*) : \ \Sigma \text{ such that } MCORD(\Sigma) \leq c^* \sqrt{\frac{\log(p)}{n}} \right\} \geq \frac{1}{7},$$

for all n and p, where the infimum is taken over all possible clustering algorithms. Multiples examples show that this $\sqrt{\log(p)/n}$ rate is attained in many different settings, including settings where

- the number K of clusters is $K = 2$ or $K = p/2$;
- the size m of the smallest cluster is $m = 2$ or $m = p/K$.

In particular, the minimax rate $\sqrt{\log(p)/n}$ still holds when we restrict to the partitions having a minimal block size m growing linearly with p.

This last result is quite surprising, since having large blocks is known to be helpful for clustering in some other contexts, as in graph clustering with Stochastic Block Model (SBM). So, can a large m help in some cases of variable clustering? While the clustering hardness looks quite insensitive to m when we look at the hardness in terms of the metric $MCORD(\Sigma)$, Bunea et al. [2] shows that a dependence in m appears when we measure the hardness in terms of the covariance gap

$$\Delta(C) = \min_{k \neq j}(C_{kk} \vee C_{jj} - C_{jk}). \tag{3}$$

We emphasize that requiring $\Delta(C) \geq \eta > 0$ is, in general, much more stringent than, requiring $MCORD(\Sigma) \geq \eta > 0$ since, when $m \geq 2$, we have

$$\Delta(C) \leq MCORD(\Sigma).$$

For example, in (2), we have $\Delta(C) = -1/2$, while $MCORD(\Sigma) = 1/2$. In particular, the covariance gap is only suited for some specific G-block covariance matrices.

For simplicity, let us illustrate the dependence in m on the simple but central case where $C = \alpha J + \tau I$, with $\alpha, \tau \geq 0$; J the $K \times K$-matrix made of ones; $\Gamma = I$; and all clusters have a size $m = p/K$. A simple SDP algorithm gives a perfect recovery of G^* with high probability, as soon as $\log(p)/n$ is small enough and

$$\Delta(C) = \tau \gtrsim \sqrt{\frac{\log(p) \vee K}{nm}} \bigvee \frac{\log(p) \vee K}{n}.$$

Furthermore, no algorithm can perfectly recover the partition with high probability when

$$\Delta(C) = \tau \lesssim \sqrt{\frac{\log(p)}{nm}} \bigvee \frac{\log(p)}{n}.$$

So, in this case, increasing m helps for clustering, at least as long as

$$m \lesssim \frac{n}{K \vee \log(p)}.$$

References

[1] F. Bunea, C. Giraud and X. Luo, *Community estimation in G-models via CORD*, arXiv:1508.01939.

[2] F. Bunea, C. Giraud, M. Royer and N. Verzelen, *A convex criterion for Latent Covariance Models*, preprint.

Distributed Statistical Estimation with Random Projections

Christina Heinze

(joint work with Brian McWilliams and Nicolai Meinshausen)

We present Loco, a communication-efficient algorithm for distributed statistical estimation. Given a matrix of features $\mathbf{X} \in \mathbb{R}^{n \times p}$ and a corresponding vector of responses, $Y \in \mathbb{R}^n$ where the dimensionality p and sample size n are very large and $p > n$, we are interested in solving the following estimation task

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} J(\boldsymbol{\beta}) := \sum_{i=1}^{n} f_i(\boldsymbol{\beta}^\top \mathbf{x}_i) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \tag{1}$$

where $\lambda > 0$ is the regularization parameter[1]. The loss functions $f_i(\boldsymbol{\beta}^\top \mathbf{x}_i)$ depend on labels $y_i \in \mathbb{R}$ and linearly on the coefficients, $\boldsymbol{\beta}$ through a vector of covariates,

[1]Throughout, $\|\cdot\|$ refers to the Euclidean norm for vectors and the spectral norm for matrices, i.e. $\|\mathbf{A}\| = \sup_{\mathbf{x}} \|\mathbf{A}\mathbf{x}\|/\|\mathbf{x}\|$.

$\mathbf{x}_i \in \mathbb{R}^p$. Furthermore, we assume all f_i to be convex and smooth with Lipschitz continuous gradients. Concretely, when $f_i(\boldsymbol{\beta}^\top \mathbf{x}_i) = (y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2$, Eq. (1) corresponds to ridge regression.

LOCO assumes that the data is distributed across K different machines (workers) (e.g. on a computing cluster) according to the features rather than the samples. This is a more challenging task for both estimation and optimization since the columns are typically assumed to have arbitrary dependencies and most commonly used loss functions are not separable over the features.

Formally, let $\mathcal{P} = \{1, \ldots, p\}$ be the set of indices. We partition this set into K non-overlapping subsets $\mathcal{P}_1, \ldots, \mathcal{P}_K$ of equal size, $\tau = p/K$ so $\mathcal{P} = \bigcup_{k=1}^K \mathcal{P}_k$ and $|\mathcal{P}_1| = |\mathcal{P}_2|, \ldots, = |\mathcal{P}_K| = \tau$.[2] A naive attempt at parallelizing (1) would simply be solving the minimization problem on each subset of features $\mathcal{P}_k$ independently. However, important dependencies between features on different workers would not in general be preserved.

We can rewrite (1) making explicit the contribution from worker k. Letting $\mathbf{X}_k \in \mathbb{R}^{n \times \tau}$ be the sub-matrix whose columns correspond to the coordinates in $\mathcal{P}_k$ (the "raw" features of worker k) and $\mathbf{X}_{(-k)} \in \mathbb{R}^{n \times (p-\tau)}$ be the remaining columns of $\mathbf{X}$, we have

$$J(\boldsymbol{\beta}) = \sum_{i=1}^n f_i \left(\mathbf{x}_{i,k}^\top \boldsymbol{\beta}_{\text{raw}} + \mathbf{x}_{i,(-k)}^\top \boldsymbol{\beta}_{(-k)} \right) + \lambda \big(\|\boldsymbol{\beta}_{\text{raw}}\|^2 + \|\boldsymbol{\beta}_{(-k)}\|^2 \big) \tag{2}$$

where $\mathbf{x}_{i,k}$ and $\mathbf{x}_{i,(-k)}$ are the rows of $\mathbf{X}_k$ and $\mathbf{X}_{(-k)}$ respectively. The idea behind our approach is to provide each worker with a low-dimensional approximation to $\mathbf{X}_{(-k)}$ which captures the contribution of these non-local features to the loss function.

We construct such an approximation using Johnson-Lindenstrauss random projections. Each worker computes a random projection of its respective block of features which we denote by $\widehat{\mathbf{X}}_k = \mathbf{X}_k \boldsymbol{\Pi}_k \in \mathbb{R}^{n \times \tau_{subs}}$. Specifically, we use the Subsampled Randomized Hadamard Transform as it allows for a computation of the matrix-vector product in $O(\tau \log \tau)$. Subsequently, each worker k constructs the matrix

$$\bar{\mathbf{X}}_k \in \mathbb{R}^{n \times (\tau + (K-1)\tau_{subs})} = \left[\mathbf{X}_k, \tilde{\mathbf{X}}_k \right], \quad \tilde{\mathbf{X}}_k = \left[\widehat{\mathbf{X}}_{k'} \right]_{k' \neq k}$$

which is the column-wise concatenation of the raw feature matrix $\mathbf{X}_k$ and the random approximations from all other workers, $\tilde{\mathbf{X}}_k$. So $\tilde{\mathbf{X}}_k \in \mathbb{R}^{n \times (K-1)\tau_{subs}}$ is the matrix whose columns are a low-dimensional approximation to $\mathbf{X}_{(-k)}$, i.e. to the columns of $\mathbf{X}$ not in $\mathbf{X}_k$, and $\tau_{subs} \ll \tau$. We shall call the columns in $\tilde{\mathbf{X}}_k$ the "random" features of worker k. This procedure is described in Figure 1.

After having constructed these local design matrices consisting of raw and random features, for a single worker the local, approximate primal problem is then

$$\min_{\bar{\boldsymbol{\beta}} \in \mathbb{R}^{\tau + (K-1)\tau_{subs}}} J_k(\bar{\boldsymbol{\beta}}) := \sum_{i=1}^n f_i(\bar{\boldsymbol{\beta}}^\top \bar{\mathbf{x}}_i) + \frac{\lambda}{2} \|\bar{\boldsymbol{\beta}}\|^2 \tag{3}$$

[2]This is for simplicity of notation only, in general the partitions can be of different sizes.

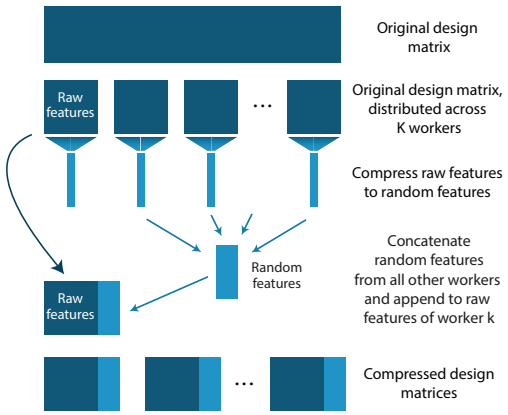

FIGURE 1. Schematic for the approximation of a large data set in a distributed fashion using random projections.

where $\bar{\mathbf{x}}_i \in \mathbb{R}^{\tau+(K-1)\tau_{subs}}$ is the i^{th} row of $\bar{\mathbf{X}}_k$. Due to computational reasons it is beneficial to solve the corresponding local dual problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\sum_{i=1}^{n} f_i^*(\alpha_i) - \frac{1}{2n\lambda}\boldsymbol{\alpha}^\top \tilde{\mathbf{K}}_k \boldsymbol{\alpha}, \quad \tilde{\mathbf{K}}_k = \bar{\mathbf{X}}_k \bar{\mathbf{X}}_k^\top. \tag{4}$$

Finally, each worker maps its local dual solution $\tilde{\boldsymbol{\alpha}}_k$ to the primal solution corresponding only to the coordinates in $\mathcal{P}_k$: $\widehat{\boldsymbol{\beta}}_k = -\frac{1}{n\lambda}\mathbf{X}_k^\top \tilde{\boldsymbol{\alpha}}_k$. In this way, each worker returns coefficients corresponding only to its own raw features. The final primal solution vector is obtained by concatenating the K local solutions.

In summary, LOCO requires only a single round of communication where low-dimensional, structured random projections are used to approximate the dependencies between features available to different workers. The structured random projections are cheap to compute and must only be communicated once.

In [1, 2], we show that LOCO has bounded approximation error with respect to the solution of Eq. (1) which only depends weakly on the number of workers. We compare LOCO against a state-of-the-art distributed optimization method on a variety of real world datasets and show that it obtains better speedups while retaining good accuracy. In particular, LOCO allows for fast cross validation as only part of the algorithm depends on the regularization parameter.

REFERENCES

[1] C. Heinze, B. McWilliams, N. Meinshausen and G. Krummenacher, *Loco: Distributing Ridge Regression with Random Projections*, arXiv preprint arXiv:1406.3469, 2014.

[2] C. Heinze, B. McWilliams and N. Meinshausen, *Dual-Loco: Distributing Statistical Estimation Using Random Projections*, to appear in AISTATS 2016.

Shape constrained estimation in low and high dimensions

John Lafferty
(joint work with Min Xu, Sabyasachi Chatterjee)

Shape constrained inference is concerned with the properties of estimators that impose geometric or structural constraints such as convexity, monotonicity, or log-concavity. Shape constraints arise naturally in certain applications such as imaging, or in economics where utility functions may be concave or monotonic. While shape constrained estimation is a classical topic, going back at least to Grenander in the 1950s, it has been gaining interest in recent years, with new problems arising under the perspective of high dimensional data analysis. In this talk we report on some recent results for shape constrained estimation in both low and high dimensions.

Nonparametric statistical theory is dominated by smoothness assumptions. But smoothness is sometimes best considered to be "wishful thinking"—this is why adaptivity properties of estimators are so important. On the other hand, shape constraints may be reasonable given knowledge of the problem. Shape constraints can also be imposed for computational reasons, as the estimators are often free of tuning parameters, and naturally lead to procedures based on convex optimization. In spite of the fundamental and classical nature of shape constrained estimation, the topic is not well understood theoretically, even in low dimensions.

In this talk we presented three recent results concerning shape constrained estimation. The first result, which appears as part of Min Xu's Ph.D. dissertation, shows that variable selection for nonparametric regression in high dimensions is qualitatively different under convexity assumptions than under smoothness assumptions, in spite of the fact that convexity and second-order smoothness have similar minimax and metric entropy properties. The second result, with Sabyasachi Chatterjee, introduces a graph-structured generalization of isotonic regression, based on the notion of a "flow." The third result, also with Sabyasachi Chatterjee, concerns adaptivity of the least squares estimator for unimodal regression.

In more detail, our results for high dimensional sparse convex regression show that variable selection using a potentially misspecified convex additive model is faithful, in the sense that there are no false negatives in the population setting. Additionally, we show that "sparsistent" variable selection is achievable using an additive model, with sample complexity scaling as $n^{4/5} \geq Cs^5\sigma^2 \log^2 p$ where s is the number of relevant variables, and p is the ambient dimension. This is interesting and perhaps surprising because under traditional smoothness assumptions, consistent variable selection in high dimensions is only possible if the sample size grows exponentially in the intrinsic dimension, $n \geq \exp(s)$ [4]. However, in the smooth setting no efficient algorithms are known, and additive models don't work in general, under model misspecification. Examples where false negatives occur are easy to construct. Thus, the geometry and shape restrictions play an essential role in high dimensional variable selection.

The second part of the talk described graph structured forms of isotonic regression. We defined the notion of a tree flow, and presented results on the statistical properties of the least squares estimator. Consider a rooted tree, and imagine a fluid flowing into the root node and dividing among the children—possibly with some leakage. The fluid is recursively divided as it flows down the tree. Thus, the flow μ_j at a node satisfies $\mu_j \geq \sum_{k \in \mathcal{C}(j)} \mu_k$ where $\mathcal{C}(j)$ is the set of children nodes. We observe a noisy flow $Y = \mu + \epsilon$ where the number of nodes is n, and ϵ denotes independent Gaussian noise. This generalizes isotonic regression, since every path from the root to a leaf is monotonic decreasing. As a motivating example, we described how profilers of computer programs measure the time or storage used in different parts of the execution tree of the program. The compiled code is instrumented to monitor the performance, but this introduces side effects that can mask the true behavior of the program. Thus, statistical profilers sample to allow the program to perform closer to its true execution behavior. Having a good handle on flow denoising could potentially allow for minimal intervention while still obtaining good estimates of the performance.

Our main result on flow denoising is primarily of interest when the depth h_n of a family of trees $\mathcal{T}_n$ grows at most logarithmically. The result says that that the risk of the least squares estimator $\widehat{\mu}_n$—which is the projection onto the convex cone of flows—satisfies the upper bound

$$\frac{1}{n}\mathbb{E}\|\widehat{\mu}_n - \mu\|^2 = \widetilde{O}\left(\frac{h_n}{n} + \frac{\mu_1\sqrt{h_n}}{n}\right) \tag{1}$$

where μ_1 is the flow at the root, with the notation $\widetilde{O}$ suppressing logarithmic factors in n. The proof is based on estimation of covering numbers for flows using a version of "Maurey's argument," together with bounds on a Gaussian supremum function in terms of Dudley's metric entropy integrals [1].

The result raises a natural question. It is well known that the risk of the least squares estimator for isotonic regression scales as $O(n^{-2/3})$, which is minimax optimal. But for logarithmic depth, the risk of the LSE for flows scales as $O(\log^3 n/n)$. Is the path the "hardest" flow estimation problem, and the star the "easiest" problem? We provide a partial answer to this question by studying a family of trees $\mathcal{T}_{n,\alpha}$ where the root has n^α children, each of which starts a single path of length $n^{1-\alpha}$, for $0 \leq \alpha \leq 1$. We give upper bounds on the least squares estimator that indicate the path and star are not, in fact, extremal. In particular, the risk of $\widehat{\mu}_n$ scales as $\widetilde{O}(n^{-1/2})$ for $\frac{1}{4} \leq \alpha \leq \frac{1}{2}$, and our simulations suggest that these bounds may be tight. However, a gap exists between our current best minimax lower bounds and our upper bounds for the LSE. This gap and other aspects of tree flows are interesting topics for future research.

References

[1] S. Chatterjee, *A new perspective on least squares under convex constraint*, *Ann. Stat.*, 42(6):2340–2381, 2014.

[2] S. Chatterjee and J. Lafferty, *Adaptive risk bounds in unimodal regression*, arXiv:1512.02956.

[3] S. Chatterjee and J. Lafferty, *Denoising flows on trees*, arXiv:1602.08048.
[4] L. Comminges and A. Dalalyan, *Tight conditions for consistency of variable selection in the context of high dimensionality*, *Ann. Stat.*, 40:2667–2696, 2012.
[5] M. Xu, M. Chen, and J. Lafferty, *Faithful variable selection in high dimensional convex regression*, arXiv:1411.1805, *Ann. Stat.* to appear.

On centrality in random growing trees: Confidence for source estimators and persistence

Po-Ling Loh

We discuss recent work regarding random growing trees. A common theme is the evolution of the most central node(s) in such trees and how they behave probabilistically as the tree grows. Our first result concerns source estimation in a diffusion spreading over a regular tree. We show that it is possible to construct confidence sets for the diffusion source with size independent of the number of infected nodes. Our estimators are motivated by analogous results in the literature concerning identification of the root node in preferential attachment and uniform attachment trees; at the core of our proofs is a probabilistic analysis of Pólya urns corresponding to the number of uninfected neighbors in specific subtrees of the infection tree. We then turn to the problem of persistence, and demonstrate that the aforementioned confidence sets have the property of persistence (i.e., settling down after finitely many steps), with probability 1. Our theory holds for regular diffusion trees, as well uniform attachment and linear and sublinear preferential attachment trees.

References

[1] J. Khim and P. Loh, *Confidence sets for the source of a diffusion in regular trees*, arXiv preprint 1510.05461, October 2015.
[2] V. Jog and P. Loh, *Persistence of centrality in random growing trees*, arXiv preprint 1511.01975, November 2015.
[3] V. Jog and P. Loh, *Analysis of centrality in sublinear preferential attachment trees via the CMJ branching process*, arXiv preprint 1601.06448, January 2016.

A Framework for Assumption-Free Predictive Regression Inference

Jing Lei
(joint work with Max G'Sell, Alessandro Rinaldo, Ryan Tibshirani, Larry Wasserman)

Consider regression data

$$Z_1, \ldots, Z_n \sim P \tag{1}$$

where $Z_i = (X_i, Y_i)$, $Y_i \in \mathbb{R}$, $X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d$ and $d \equiv d_n$ is allowed to increase as n increases. Let

$$\mu(x) = \mathbb{E}[Y|X = x] \tag{2}$$

denote the regression function. We are interested in predicting a new Y from a new X with no assumptions on $\mu(x)$, P or the design matrix. The main goal of this paper is to construct a prediction set $C \subset \mathbb{R}^d \times \mathbb{R}$, such that $\mathbb{P}(Y \in C(X)) \geq 1-\alpha$, where $C(x) = \{y : (x, y) \in C\}$. A secondary goal is to construct model-free inferential statements about the importance of each covariate.

Our leading example is high-dimensional regression where $d \gg n$ and a linear function is used to approximate $\mu(\cdot)$ but is not assumed to be correct. Common approaches to high-dimensional linear regression include the lasso, forward stepwise selection and the elastic net. But there is very little work on prediction sets.

We construct coverages sets for the response Y_{n+1} at a future predictor X_{n+1}. The resulting prediction set inherits the good theoretical properties of the original estimator under standard assumptions, while maintaining finite sample validity under essentially no assumptions. The basis of our approach is *conformal prediction*, a method invented by [1]. The conformal approach was further developed in [2, 4, 3].

Conformal prediction intervals naturally avoids overfitting, and, somewhat remarkably, are guaranteed to deliver proper coverage in finite sample without any assumptions on P or on $\hat{\mu}$. The main feature of the conformal prediction procedure is the *symmetry* between the n data points $Z_1, \ldots, Z_n$ and the new data point X_{n+1} at which prediction is wanted for Y_{n+1}.

Consider the following strategy: for each value $y \in \mathbb{R}$, we construct an augmented regression estimate $\hat{\mu}_y$, which is trained on the augmented data set $Z_1, \ldots, Z_n$, (X_{n+1}, y). Now we define

$$R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|,\ i = 1, \ldots, n \quad \text{and} \quad R_{y,n+1} = |y - \hat{\mu}_y(X_{n+1})|, \tag{3}$$

and we rank $R_{y,n+1}$ among the remaining fitted residuals $R_{y,1}, \ldots, R_{y,n}$, computing

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(R_{y,i} \leq R_{y,n+1}) = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} I(R_{y,i} \leq R_{y,n+1}), \tag{4}$$

the proportion of points in the augmented sample whose fitted residual is smaller than the last one, $R_{y,n+1}$. By symmetry, when evaluated at $y = Y_{n+1}$, we see that the constructed statistic $\pi(Y_{n+1})$ is uniformly distributed over the set $\{1/(n+1), 2/(n+1), \ldots, 1\}$, which implies

$$\mathbb{P}\Big((n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil\Big) \geq 1 - \alpha. \tag{5}$$

We may interpret the above property as saying that $\pi(Y_{n+1})$ provides a valid (conservative) p-value for the test $H_0 : Y_{n+1} = y$. Furthermore, the property (5) immediately leads to our conformal prediction interval at X_{n+1}, namely

$$C(X_{n+1}) = \Big\{y \in \mathbb{R} : (n+1)\pi(y) \leq \lceil (1-\alpha)(n+1) \rceil\Big\}. \tag{6}$$

The steps in (3), (4), (6) must be repeated each time we want to produce a prediction interval (at a new feature value). Also, in practice, we must restrict our attention in (6) to a discrete grid of trial values y.

Theorem 0.3. *If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. pair (X_{n+1}, Y_{n+1}),*

$$\mathbb{P}\Big(Y_{n+1} \in C(X_{n+1})\Big) \geq 1 - \alpha,$$

for the conformal prediction band $C^{\text{conf}}_{1-\alpha}$. If we assume additionally that for all $y \in \mathbb{R}$, the fitted absolute residuals $R_{y,i} = |Y_i - \hat{\mu}_y(X_i)|$, $i = 1, \ldots, n$ have a continuous joint distribution, then it also holds that

$$\mathbb{P}\Big(Y_{n+1} \in C(X_{n+1})\Big) \leq 1 - \alpha + \frac{1}{n+1}.$$

No assumptions are needed in this theorem about the the regression estimator $\hat{\mu}$. The validity and accuracy of the conformal interval holds over all P and all $\hat{\mu}$. This is a somewhat remarkable and unique property of conformal inference. Generally speaking, as we improve our estimate $\hat{\mu}$ of the underlying regression function μ, the resulting conformal prediction interval decreases in length. Intuitively, this happens because a more accurate $\hat{\mu}$ leads to a more accurate estimate of the residual distribution, and conformal intervals are essentially defined by the quantiles of the (augmented) residual distribution.

There is an alternative approach, called "split conformal prediction" — called *inductive conformal inference* in [5, 1] — that significantly reduces the computational cost of conformal inference. Split conformal prediction separates the fitting and ranking steps in conformal prediction using sample splitting. This method eliminates the need to do the fitting over all (x, y). To be specific, the input data set is split into two subsets. The first subsample is used to fit the regression function $\hat{\mu}$. For a new data X_{n+1}, the conformal prediction interval includes all y such that $|y - \hat{\mu}(X_{n+1})|$ ranks no higher than $(n/2 + 1)(1 - \alpha)$ among all residuals in the second sample.

Theorem 0.4. *The split conformal algorithm satisfies, for any P,*

$$\mathbb{P}(Y \in C_{\text{split}}(X)) \geq 1 - \alpha. \tag{7}$$

If the residuals have continuous joint distribution then

$$1 - \alpha \leq \mathbb{P}(Y \in C_{\text{split}}(X)) \leq 1 - \alpha + \frac{2}{n+2}.$$

The full conformal and split conformal methods both tend to produce prediction bands $C(x)$ whose width (i.e., length, we will use these two terms interchangeably) is roughly constant over $x \in \mathbb{R}^d$. In some scenarios this will not be true, i.e., the residual variance will vary nontrivially with X, and we will want the conformal band to adapt correspondingly.

We now introduce an extension to the conformal method that can account for non-constant residual variance. Our idea is based estimating the mean absolute deviation (MAD) of the fitted residual rather than the standard deviation, since

the former exists in some cases in which the latter does not. Recall that, in order for the conformal inference method to have valid coverage, we can actually use any conformity score function to generalize the definition of (absolute) residuals. Then, for the present extension, we simply modify the definition of residuals to use *locally-weighted* residuals

$$R_{y,i} = \frac{|Y_i - \hat{\mu}_y(X_i)|}{\hat{\rho}_y(X_i)},\ i = 1, \ldots, n, \quad \text{and} \quad R_{y,n+1} = \frac{|y - \hat{\mu}_y(x)|}{\hat{\rho}_y(x)}, \tag{8}$$

where now $\hat{\rho}_y(x)$ denotes an estimate of the conditional MAD of $(Y - \mu(X))|X = x$, as a function of $x \in \mathbb{R}^d$. With the locally-weighted residuals in (8), the validity and accuracy properties of the full conformal inference method carry over.

References

[1] Vovk, V., Gammerman, A., & Shafer, G. *Algorithmic Learning in a Random World* (2005), Springer.

[2] Lei, J., Robins, J., & Wasserman, L. *Distribution free prediction sets*, Journal of the American Statistical Association **108** (2013), 278–287.

[3] Lei, J., & Wasserman, L. *Distribution-free prediction bands for non-parametric regression*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** (2014), 71–96.

[4] Lei, J., Rinaldo, A., & Wasserman, L. *A conformal prediction approach to explore functional data*, Annals of Mathematics and Artificial Intelligence **74** (2015), 29–43.

[5] Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). *Inductive confidence machines for regression*, In *Machine Learning: ECML 2002* (2002), 345–356. Springer.

Statistical Blind Source Separation - with Applications in Cancer Genetics

Axel Munk

(joint work with Merle Behr (Göttingen), Chris Holmes (Oxford))

The SBSSR model.

We are studying a particular kind of blind source separation problem embedded in a change-point regression setting. In blind source separation problems one observes a mixture of source functions, and aims to recover the original sources from the available observations. The blindness refers to the fact that neither the sources nor the mixing is known. We consider single linear mixtures of step functions with a known finite alphabet in a Statistical Blind Source Separation Regression (SBSSR) model.

More precisely, for a given finite alphabet $\mathfrak{A} \subset \mathbb{R}$, a given number of source components $m \geq 2$, unknown step functions $f = (f^1, \ldots, f^m)^\top$ each taking values in the known alphabet, $\mathrm{imag}(f^i) \subset \mathfrak{A}$, and unknown probability mixing weights $\omega = (\omega_1, \ldots, \omega_m)^\top \in \mathbb{R}^m_+$ with $\sum_{i=1}^m \omega_i = 1$, one observes from the mixture

$$Y_j = \omega^\top f(x_j) + \epsilon_j = \sum_{i=1}^m \omega_i f^i(x_j) + \epsilon_j, \quad j = 1, \ldots, n, \tag{1}$$

where ϵ is normal noise with mean zero. We assume that $\omega^\top f$ in (1) is sampled equidistantly at $x_j := (j-1)/n$, $j = 1, \ldots, n$ and that all step functions f^i are defined on the domain $[0, 1)$. Extensions to more general domains $\subset \mathbb{R}$ and sampling designs are straightforward under suitable assumptions. The aim in model (1) is to estimate ω and f from the observations $Y = (Y_1, \ldots, Y_n)$ and to construct honest confidence statements for all quantities.

Identifiability.

We stress that already in the noiseless case, i.e., $\epsilon \equiv 0$ in (1), it is far from obvious under which criteria the weights ω and the sources f are identifiable. We characterize the identifiability issue as a combinatorial problem and derive simple sufficient and necessary identifiability criteria which, to the best of our knowledge, has been elusive. On the one hand, these conditions ensure discriminability of different mixture values, which is a necessary conditions on ω to guarantee identifiabilty of f, and, on the other hand, they ensure a sufficient variability of f, which is necessary to guarantee identifiabilty of ω.

Moreover, we discuss how likely it is for the derived identifiability criteria to be satisfied when the mixing weights are drawn from the uniform distribution and when the underlying sources are discrete Markov processes. We show that the mixture becomes identifiable exponentially fast, which reveals identifiability not to be an issue in most practical situations. See [1] for more details on the identifiability issue in model (1).

The SESAME estimation methodology.

In the regression setting we propose a new methodology, called SESAME (SEparateS finite Alphabet MixturEs), which yields uniform confidence sets and optimal estimation rates (up to log-factors) for all parameters in model (1) under very weak identifiability conditions [2].

First, SESAME provides honest confidence regions $\mathcal{C}_{1-\alpha}(Y)$ for the mixing weights ω which are characterized by the acceptance region of a certain multiscale test [3] with level α. Then, it estimates $\hat{\omega} \in \mathcal{C}_{1-\alpha}(Y)$, where now the level α can be seen as a tuning parameter for which we propose data driven selection methods. For $\hat{\omega}$ and $\mathcal{C}_{1-\alpha}(Y)$ we obtain almost optimal estimation rates and diameter $\ln(n)/\sqrt{n}$, respectively. Second, SESAME estimates the sources f as a constrained maximum likelihood estimator, where the constraint comes from the same multiscale statistic as for $\mathcal{C}_{1-\alpha}(Y)$ but with a possibly different level β. This yields asymptotically honest multivariate confidence bands $\mathcal{H}(\beta)$ for the sources f. For the resulting estimate $\hat{f}$ we derive exact recovery, i.e., the number of change-points of f and its function values are estimated exactly and its change-point locations with the minimax rate $1/n$ up to a log square term with probability converging to one at a superpolynomial rate.

SESAME's estimates and confidence statements can be computed efficiently using dynamic programming and certain pruning steps.

Applications.

Model (1) arises in many different applications, for instance in digital communication with mixtures of multi-level PAM signals [4]. Our motivation, however,

comes from an application in cancer genetics. We use SESAME to analyze genetic sequencing data in order to estimate clonal proportions in a tumor, and the corresponding copy number variations [5].

Acknowledgement.
This work has been initiated through intensive discussions at the MFO workshop 'Frontiers in Nonparametric Statistics' in 2012.

References

[1] M. Behr and A. Munk. *Identifiability for blind source separation of multiple finite alphabet linear mixtures.* arXiv preprint. arXiv:1505.05272.
[2] M. Behr, C. Holmes, and A. Munk. *Multiscale blind source separation.* preprint 2015.
[3] K. Frick, A. Munk, and H. Sieling. *Multiscale change point inference.* Journal of the Royal Statistical Society: Series B (Statistical Methodology). (2014) **76** 495 – 580.
[4] J. G. Proakis. *Digital communications.* Communications and Signal Processing. (1995) McGraw-Hill
[5] B. Liu, C.D. Morrison, C.S. Johnson, D.L. Trump, M. Qin, J.C. Conroy, J. Wang, and S. Liu. *Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges.* Oncotarget. (2013) **4** 1868.

Invariances and Causality

Jonas Peters
(joint work with Peter Bühlmann, Nicolai Meinshausen)

The detection of statistical dependences is a core problem of statistics. In many situations, however, we prefer a **causal model** over a purely predictive one since the former can also predict what happens under interventions. For example, there might be a correlation between a certain disease and drinking wine but only a causal model tells whether changing one's drinking behaviour prevents us from getting ill.

In order to formulate and tackle those causal problems, we use the language of **structural equation models** (SEMs) [9]. In SEMs, each variable $X_j, j = 1, \dots, p$ is modeled as a deterministic function of its direct causes $X_{pa(j)}$ and some noise variable N_j, that is

$$X_j = f_j(X_{\mathbf{PA}_j}, N_j), \quad j = 1, \dots, p,$$

where all noise variables are assumed to be jointly independent. SEMs do not only allow us to model observational distributions; at the same time we can also use them in order to model what happens under interventions, i.e., when some of the variables are actively set to specific values (e.g. gene knockouts or randomized studies).

In **causal discovery** (or structure learning), we try to learn the causal structure from observational and/or interventional data. Using the concept of SEMs, it becomes apparent that without further assumptions, this goal is impossible to achieve: any observational distribution can be modeled by several SEMs with different graphs. Under further assumptions, however, this problem becomes solvable. Under faithfulness [e.g. 15], for example, the Markov equivalence class of the

underlying graph becomes identifiable [17, 1, 16]. Spirtes et al. [15], Chickering [3], Castelo and Kocka [2], Kalisch and Bühlmann [8], He and Geng. [5], Hauser and Bühlmann [4], (and others) provide methods that infer the identifiable structure from data. More recently, work has been done for fully identifiable structures exploiting additional restrictions such as non-Gaussianity [14], nonlinearity [6, 11] or equal error variances [10]. Janzing et al. [7] exploit an independence between causal mechanisms.

In this talk, we discuss an approach to causal discovery that is called **invariant causal prediction** [12]. In many situations, we are interested in the system's behavior under a change of environment. Here, causal models become important because they are usually considered invariant under those changes. A causal prediction (which uses only direct causes of the target variable as predictors) remains valid even if we intervene on predictor variables or change the whole experimental setting. In this approach, we exploit invariant prediction for causal inference: given data from different experimental settings, we use invariant models to estimate the set of causal predictors for a given target variable. More formally, we consider a target variable Y and make the following assumption: There exists a set $S^* \subseteq \{1, \dots, d\}$ and $\gamma^* = (\gamma_1^*, \dots, \gamma_d^*)^t$ with support S^* that satisfies

$$\begin{aligned} &\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and} \\ &\qquad\qquad\qquad\quad Y^e = X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}\,. \end{aligned} \tag{1}$$

In the case of structural equation models [9] or potential outcomes [13], where the environments $e \in \mathcal{E}$ correspond to interventions that do not act on Y, the set $S^* := \mathbf{PA}_Y$ satisfies this assumption [12].

Our goal is to estimate S^*, which is not possible in many situations. Instead, we now construct a surrogate $S(\mathcal{E})$ and discuss its relation to S^* below. We therefore introduce the following null hypothesis:

$$H_{0,S}(\mathcal{E}): \begin{cases} \text{for all } e \in \mathcal{E}: & X^e \text{ has an arbitrary distribution and} \\ & Y^e = X^e\beta^{pred,e}(S) + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*} \end{cases},$$

where

$$\beta^{pred,e}(S) := \operatorname{argmin}_{\beta \in \mathbb{R}^p : \beta_k = 0 \text{ if } k \notin S} E(Y^e - X^e\beta)^2$$

are the least-squares population regression coefficients. We are now able to define

$$S(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S}(\mathcal{E}) \text{ is true}} S,$$

which can be estimated by

$$\hat{S}(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S}(\mathcal{E}) \text{ not rejected}} S$$

where ideas for statistical tests can be found in [12]. Given that (1) is satisfied for S^*, we have the following statements:

i) $S(\mathcal{E}) \subseteq S^*$,
ii) $\mathcal{E} = 1 \Rightarrow S(\mathcal{E}) = \emptyset$,

iii) $\mathcal{E}_1 \supseteq \mathcal{E}_2 \Rightarrow S(\mathcal{E}_1) \supseteq S(\mathcal{E}_2)$ and
iv) $\mathbb{P}(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha$.

Robustness properties, ideas for extending the framework to hidden variables, results for artificial and real data sets are discussed in [12].

References

[1] S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

[2] R. Castelo and T. Kocka. On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research*, 4:527–574, 2003.

[3] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

[4] A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318, 2015.

[5] Y.-B. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

[6] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 689–696, 2009.

[7] D. Janzing, J. M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

[8] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

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

[10] J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.

[11] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[12] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (to appear, with discussion)*, 2016.

[13] D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331, 2005.

[14] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[15] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, USA, 2nd edition, 2000.

[16] J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–522, 2001.

[17] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.

Fast rates for TV denoising

Philippe Rigollet
(joint work with Jan-Christian Hütter)

Motivated by its practical success, we show that the two-dimensional total variation denoiser [ROF92] satisfies a sharp oracle inequality that leads to near optimal rates of estimation for a large class of image models such as bi-isotonic, Hölder smooth and cartoons.

Consider a noisy image $y \in \mathbb{R}^{N\times N}$, which we will identify with a vector $y \in \mathbb{R}^n$, $n = N^2$, and the $N \times N$ grid graph defined on the vertex set $[N]^2$ which contains edge $e = \big([i,j],[k,l]\big)$ in its edge set E if and only if $[k,l]-[i,j] \in \{[1,0],[0,1]\}$. The total variation (TV) denoiser is defined as

$$\hat\theta \in \operatorname*{argmin}_{\theta\in\mathbb{R}^n} \frac{1}{n}\|\theta - y\|_2^2 + \lambda\|D\theta\|_1\,, \tag{1}$$

where D denotes the incidence matrix of the 2D grid graph and $\lambda > 0$ is a tuning parameter. The TV denoiser is often used to restore blurry images due to its reported properties of smoothing out grainy regions while allowing for sharp boundaries between regions of different signal intensity. Being a convex program, it can be solved efficiently, see [AT16] and the references therein.

Consider the Gaussian sequence model

$$y = \theta^* + \varepsilon\,, \tag{2}$$

where $\theta^* \in \mathbb{R}^n$ is the unknown parameter of interest and $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_n)$ is an isotropic Gaussian random vector. Most of the previous analysis of the performance of the TV denoiser has been focused on sparsistency results, i.e. recovering the places where θ^* has a jump, see [QJ12, SSR12, OV15, VLLHP16]. Here however, we are interested in the averaged performance described the mean squared error $\|\hat\theta - \theta^*\|_2^2/n$. Previous works investigating this are [DHL14, NW13] and [WSST15], which already contains a $n^{-4/5}$ rate for mean squared error in the 2D case. The main theorem below improves these results to a fast n^{-1} rate, up to logarithmic factors, and allows for model misspecification.

Theorem 1.5. *Fix $\delta \in (0,1)$. Then there exist constants $C, c > 0$ such that the TV denoiser $\hat\theta$ defined in (1) with $\lambda = c\sigma\sqrt{(\log n)\log(en/\delta)}/n$ satisfies*

$$\begin{aligned}\frac{1}{n}\|\hat\theta - \theta^*\|^2 \le \inf_{\substack{\bar\theta\in\mathbb{R}^n\\ T\subseteq E}} \left\{\frac{1}{n}\|\bar\theta - \theta^*\|^2 + 4\lambda\|(D\bar\theta)_{T^c}\|_1\right\}\\ + \frac{C\sigma^2}{n}\left(|T|(\log n)\log(en/\delta) + \log(e/\delta)\right),\end{aligned} \tag{3}$$

with probability $1-\delta$, where $(D\bar\theta)_{T^c}$ denotes the restriction of $D\bar\theta$ to the subset $T^c \subset E$. In particular, it yields

$$\frac{1}{n}\|\hat\theta - \theta^*\|_2^2 \lesssim \frac{\sigma\|D\theta^*\|_1 \wedge \sigma^2\|D\theta^*\|_0}{n}\log^2(en/\delta)$$

with $\|D\theta^*\|_0$ being the number of nonzero components of $D\theta^*$.

The key novelty of our proof is a sharp control of the maximum column norm of the pseudo-inverse $D^\dagger$. Specifically, we use the representation $D^\dagger = (D^\top D)^\dagger D^\top$ together with the spectral decomposition of the graph Laplacian $D^\top D$. While these results mostly pertain to the 2D grid, they can be extended to other graphs, such as the complete graph or random graphs.

Having established Theorem 1.5, we can use the trade-off in (3) to get almost minimax rates for certain function classes on the 2D grid by setting $\bar{\theta}$ to a linear approximation of $\hat{\theta}$. For α-Hölder functions, we have (up to log factors) $n^{-2\alpha/(2\alpha+1)}$, and $n^{-1/2}$ for piecewise constant functions as well as for the class of bi-isotonic matrices, which recently has been studied in [CGS15, Bel15].

Acknowledgments. This work was supported in part by the National Science Foundation (DMS-1317308, CAREER-DMS-1053987).

References

[AT16] Taylor B. Arnold and Ryan J. Tibshirani, *Efficient implementations of the generalized lasso dual path algorithm*, Journal of Computational and Graphical Statistics **25** (2016), no. 1, 1–27.

[Bel15] Pierre C. Bellec, *Sharp oracle inequalities for Least Squares estimators in shape restricted regression*, arXiv preprint arXiv:1510.08029 (2015).

[CGS15] Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen, *On matrix estimation under monotonicity constraints*, arXiv preprint arXiv:1506.03430 (2015).

[DHL14] Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer, *On the prediction performance of the lasso*, to appear in Bernoulli, arXiv 1402.1700, February 2014.

[NW13] Deanna Needell and Rachel Ward, *Near-optimal compressed sensing guarantees for total variation minimization*, IEEE Transactions on Image Processing **22** (2013), no. 10, 3941–3949.

[OV15] Edouard Ollier and Vivian Viallon, *Regression modeling on stratified data: automatic and covariate-specific selection of the reference stratum with simple L_1-norm penalties*, arXiv:1508.05476 [math, stat] (2015).

[QJ12] Junyang Qian and Jinzhu Jia, *On pattern recovery of the fused Lasso*, arXiv:1211.5194 (2012).

[ROF92] Leonid I. Rudin, Stanley Osher, and Emad Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1, 259–268.

[SSR12] James Sharpnack, Aarti Singh, and Alessandro Rinaldo, *Sparsistency of the edge lasso over graphs*, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12) (Neil D. Lawrence and Mark A. Girolami, eds.), vol. 22, 2012, pp. 1028–1036.

[VLLHP16] Vivian Viallon, Sophie Lambert-Lacroix, Hölger Hoefling, and Franck Picard, *On the robustness of the generalized fused lasso to prior specifications*, Statistics and Computing **26** (2016), no. 1-2, 285–301.

[WSST15] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan Tibshirani, *Trend Filtering on Graphs*, Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, 2015, pp. 1042–1050.

Spectral analysis of high-dimensional sample covariance matrices in the missing-at-random scenario

Angelika Rohde
(joint work with Kamil Jurczak)

We study asymptotic spectral properties of high-dimensional sample covariance matrices with missing observations. Let

$$Y = (Y_1, ..., Y_n) \in \mathbb{R}^{d\times n},\ Y_k = (Y_{1k}, ..., Y_{dk})^* \in \mathbb{R}^d,\ k = 1, ..., n,$$

be a sample of independent identically distributed (iid) random vectors with covariance matrix

$$T = \mathbb{E}\big((Y_1 - \mathbb{E}Y_1) \otimes (Y_1 - \mathbb{E}Y_1)\big).$$

In examples as described above, we do not observe the whole random vector Y_k but some of its components. This missingness is represented by a random matrix $\varepsilon \in \mathbb{R}^{d\times n}$ with entries

$$\varepsilon_{ik} = \begin{cases} 1 & \text{if } Y_{ik} \text{ is observed} \\ 0 & \text{if } Y_{ik} \text{ is missing.} \end{cases}$$

Under the assumption that the matrices Y and ε are independent, the estimator

$$\hat{T}_{ij} = \frac{1}{N_{ij}} \sum_{k\in\mathcal{N}_{ij}} \big(Y_{ik} - \bar{Y}_i\big)\big(Y_{jk} - \bar{Y}_j\big)$$

is the analogue of the sample covariance and hence the natural estimator for T_{ij}, where

$$\mathcal{N}_{ij} = \Big\{k \in \{1, \dots, n\} :\ \varepsilon_{ik}\varepsilon_{jk} = 1\Big\},\quad N_{ij} = 1 \vee \#\mathcal{N}_{ij} \tag{1}$$

and

$$\bar{Y}_i = \frac{1}{N_{ii}} \sum_{k\in\mathcal{N}_{ii}} Y_{ik}.$$

Subsequently, $\hat{T} = (\hat{T}_{ij}) \in \mathbb{R}^{d\times d}$ is referred to as sample covariance matrix with missing observations. If $\mathbb{E}Y_k = 0$ is known in advance one typically uses the estimator

$$\hat{\Sigma} = \big(\hat{\Sigma}_{ij}\big) \in \mathbb{R}^{d\times d},\quad \hat{\Sigma}_{ij} = \frac{1}{N_{ij}} \sum_{k\in\mathcal{N}_{ij}} Y_{ik}Y_{jk}.$$

In what follows we write $\hat{\Xi}$ for $\hat{T}$ and $\hat{\Sigma}$ if a statement holds for both estimators.

1. Assumptions

Let $(X(i,k))_{i,k\in\mathbb{N}}$ be a double array of iid centered random variables with unit variance. The left upper $d \times n$ submatrix is denoted by $X_{d,n}$. Then the random vectors $Y_{1,d,n}, \dots, Y_{n,d,n} \in \mathbb{R}^d$ are the columns of the matrix

$$Y_{d,n} - \mathbb{E}Y_{d,n} = T_{d,n}^{1/2} X_{d,n}.$$

with

$$T_{d,n} = \operatorname{diag}(T_{11,d,n}, \dots, T_{dd,d,n}) \in \mathbb{R}^{d\times d}.$$

This structure on the population covariance matrix is the simplest one which allows to visualize the effects of missing observations on the spectrum of the sample covariance matrix. In this article we investigate asymptotic spectral properties of $\hat{\Xi}$ under the classical missing (completely) at random (MAR) setting. $(\varepsilon_{d,n})_{d,n}$ is a triangular array of random matrices $\varepsilon_{d,n} \in \mathbb{R}^{d\times n}$ independent of $(X(i,k))_{i,k\in\mathbb{N}}$, where the entries $\varepsilon_{ik,d,n}$ are independent Bernoulli variables with observation probabilities

$$\mathbb{P}(\varepsilon_{ik,d,n} = 1) = p_{i,d,n},\ i = 1,\dots,d,\ k = 1,\dots n.$$

The dependence of the set $\mathcal{N}_{ij}$ and the number N_{ij} in (1) on the sequence $(\varepsilon_{d,n})$ is indicated by an additional subscript d,n. Throughout this report we impose that the family of spectral measures of the population covariance matrices $(T_{d,n})$ as well as the family of empirical distributions

$$(\mu^{w_{d,n}})_{d,n}, \quad \text{with } \mu^{w_{d,n}} = \frac{1}{d}\sum_{i=1}^{d}\delta_{w_{i,d,n}} \text{ and } w_{d,n} = \left(p_{1,d,n}^{-1},...,p_{d,d,n}^{-1}\right),$$

are tight. Asymptotic statements refer to $d \to \infty$ while $n = n(d)$ satisfies $\limsup_{d\to\infty}(d/n) < \infty$. The sequence of sample covariance matrices with missing observations is denoted by $(\hat{\Xi}_{d,n})_{d,n}$, the corresponding sequence of spectral measures by $(\mu_{d,n})_{d,n}$ and their Stieltjes transforms by $(m_{d,n})_{d,n}$.

2. Results

Define

$$S_{d,n} = \operatorname{diag}\left(\frac{1-p_{1,d,n}}{p_{1,d,n}}T_{11,d,n},\dots,\frac{1-p_{d,d,n}}{p_{d,d,n}}T_{dd,d,n}\right)$$
$$\text{and } R_{d,n} = \operatorname{diag}\left(\frac{1}{p_{1,d,n}}T_{11,d,n},\dots,\frac{1}{p_{d,d,n}}T_{dd,d,n}\right).$$

Theorem 2.1. *Suppose that the assumptions stated in Section 1 hold, and*

$$\sup_d \|R_{d,n}\|_{S_\infty} < \infty.$$

Then for any $z \in \mathbb{C}^+$, we have $|m_{d,n}(z) - m^\circ_{d,n}(z)| \to 0$ a.s., where $m^\circ_{d,n}(z)$ satisfies

$$m^\circ_{d,n}(z) = \frac{1}{d}\operatorname{tr}\left\{\left(\frac{1}{1+\frac{d}{n}e^\circ_{d,n}(z)}R_{d,n} - S_{d,n} - zI_{d\times d}\right)^{-1}\right\}$$

and $e^\circ_{d,n}$ is the (unique) solution of the fixed point equation

$$e^\circ_{d,n}(z) = \frac{1}{d}\operatorname{tr}\left\{R_{d,n}\left(\frac{1}{1+\frac{d}{n}e^\circ_{d,n}(z)}R_{d,n} - S_{d,n} - zI_{d\times d}\right)^{-1}\right\}.$$

Moreover, $m^\circ_{d,n}$ is the Stieltjes transform of a probability measure $\mu^\circ_{d,n}$ on $\mathbb{R}$ and

$$\mu^\circ_{d,n} - \mu_{d,n} \Longrightarrow 0 \quad a.s.$$

It is well-known that the Stieltjes transform of the Marčenko-Pastur law with parameters $\left(y, \sigma^2/p_0\right)$ is the unique solution to

$$s(z) = \left(\frac{\sigma^2}{p_0} \cdot \frac{1}{1 + \frac{\sigma^2}{p_0} y s(z)} - z \right)^{-1}$$

from $\mathbb{C}^+ \to \mathbb{C}^+$. In the special case $T_{d,n} = \sigma^2 I_{d\times d}$ and $p_{d,n} = (p_0, \ldots, p_0) \in (0,1)^d$, we have

$$m^{\circ}_{d,n}\left(z - \sigma^2 \frac{1-p_0}{p_0}\right) = \left(\frac{\sigma^2}{p_0} \frac{1}{1 + \frac{d}{n}\frac{\sigma^2}{p_0} m^{\circ}_{d,n}\left(z - \sigma^2 \frac{1-p_0}{p_0}\right)} - z \right)^{-1}.$$

Hence, $\mu^{\circ}_{d,n}$ is the Marčenko-Pastur law $\mu^{MP}_{\frac{d}{n}, \frac{\sigma^2}{p_0}}$ shifted by $\sigma^2 \frac{1-p_0}{p_0}$ to the left.

Corollary 2.2. *Grant the conditions of Theorem 2.1. If $p_{i,d,n} = p_0 > 0$ for $i = 1, \ldots, d$ and $d, n \in \mathbb{N}$ and $T_{d,n} = \sigma^2 I_{d\times d}$, $\sigma^2 > 0$, we obtain*

$$\mu_{d,n} \Longrightarrow \mu^{MP}_{y, \frac{\sigma^2}{p_0}} \star \delta_{-\frac{1-p_0}{p_0}\sigma^2} \quad a.s.$$

as $d \to \infty$ and $d/n \to y > 0$.

For $\hat{\Sigma}_{d,n}$ we even determine the a.s. limit of the extremal eigenvalues.

Theorem 2.3. *Grant the conditions of Corollary 2.2 let additionally $\mathbb{E}X_{11}^4 < \infty$ and $\varepsilon_{d,n} \in \mathbb{R}^{d\times n}$ be the upper left corner of a double array $(\varepsilon(i,k))_{i,k\in\mathbb{N}}$ of iid Bernoulli variables with parameter p_0. Assume that $\mathbb{E}Y_{d,n} = 0$. Then, if $0 < y < 1$,*

$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) = \frac{\sigma^2}{p_0}\left(1 - \sqrt{y}\right)^2 - \frac{1-p_0}{p_0}\sigma^2 \quad a.s., \quad \text{and}$$
$$\lim_{d\to\infty} \lambda_{\max}\left(\hat{\Sigma}_{d,n}\right) = \frac{\sigma^2}{p_0}\left(1 + \sqrt{y}\right)^2 - \frac{1-p_0}{p_0}\sigma^2 \quad a.s.$$

The characterization of positive definiteness in the null case under the missing at random scenario is an immediate corollary of Theorem 2.3.

Corollary 2.4. *Under the condition of Theorem 2.3,*

$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) < 0 \quad a.s. \text{ if } \quad p_0 < 1 - (1 - \sqrt{y})^2, \quad \text{and}$$
$$\lim_{d\to\infty} \lambda_{\min}\left(\hat{\Sigma}_{d,n}\right) > 0 \quad a.s. \text{ if } \quad p_0 > 1 - (1 - \sqrt{y})^2.$$

Causal inference in partially linear structural equation models – identifiability and estimation

Dominik Rothenhäusler
(joint work with Jan Ernest, Peter Bühlmann)

The talk was concerned with causal inference in partially linear additive structural equation models (SEMs) with Gaussian noise. Current research covers two special cases: Recently, it has been shown that under causal minimality and if all functions in the structural equation model are nonlinear, the SEM is identifiable given only observational data [2]. On the other hand, under faithfulness and if all functions in the structural equation model are linear, the distribution equivalence class is equal to the Markov equivalence class. Consequently, there exist well-known graphical and transformational characterizations of the distribution equivalence class, see e.g. [1] and references therein. However, the intermediate case is only poorly understood.

We provide comprehensive characterizations of the distribution equivalence class in the case where some functions in the SEM may be linear and some may be not. These characterizations can be formulated from the perspective of causal orders and from a functional viewpoint. Under faithfulness, they give rise to a graphical representation of the distribution equivalence class and a transformational characterization based on covered linear edge reversals. These results comprise the aforementioned known results for solely linear and solely nonlinear SEMs.

These characterizations are leveraged in an algorithm that, given one member of the distribution equivalence class, computes a graphical representation of all DAGs in the distribution equivalence class. We prove its (high-dimensional) consistency and demonstrate its performance in simulations.

References

[1] P. Spirtes, C. N. Glymour and R. Scheines, *Causation, prediction, and search*, MIT press (2000).

[2] J. Peters, J. M. Mooij, D. Janzing and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, Journal of Machine Learning Research, **15** (2014), 2009–2053.

Increasing-domain asymptotics for inversion-free estimator of the Gaussian random fields

Clayton Scott
(joint work with Hossein Keshavarz, XuanLong Nguyen)

Abstract

Gaussian random fields are a powerful tool for modeling environmental processes. For high dimensional samples, classical approaches for estimating the covariance parameters require highly challenging and massive computations, such as the evaluation of the Cholesky factorization or solving linear systems. Recently, Anitescu, Chen and Stein [1] proposed a fast and scalable algorithm which does not need such burdensome computations. The main focus of this article is to study the asymptotic behavior of the algorithm of Anitescu et al. (ACS) for regular and irregular grids in the increasing domain setting. Despite the fact that ACS's method entails a non-concave maximization, our results hold for any stationary point of the objective function .

1. Introduction

Gaussian process (GP) models gained widespread popularity in spatial statistics and machine learning due to the versatility of their mean and covariance structure. In spatial statistics, the unknown covariance function of GP is commonly assumed to belong to a finite dimensional parametric family and its parameters must be estimated from available data. The traditional algorithms of estimating covariance parameters such as maximum likelihood estimation (MLE) and Bayesian inference can be prohibitive for large and irregularly spaced data. The main computational burden of such algorithms is directly related to solving large *system of linear equations (SLE)* which is inevitable for evaluating the log-likelihood function and its derivatives. Despite the recent advances toward scalable solution of SLEs such as iterative *Krylov subspace*, computing the MLE is still a challenging task, especially for processes sampled at numerous and irregularly spaced sites.

The computational barriers of optimizing likelihood based loss functions (e.g. full or tapered MLE) accentuate using losses whose evaluation does not require extensive computations such as inverting the covariance matrix. The first attempt toward such a goal has been done by Anitescu, Chen and Stein [1]. Their proposed loss function which is independent of the precision matrix will be referred as *ACS's* algorithm. Simulation studies verify the efficiency of ACS's method in the case that the covariance matrix has a bounded condition number. The main purpose of this paper is to study the asymptotic properties of the ACS's algorithm such

This research is partially supported by NSF grant ACI-1047871. Additionally, CS is partially supported by NSF grants 1422157, 1217880, and 0953135, and LN by NSF CAREER award DMS-1351362, NSF CNS-1409303, and NSF CCF-1115769. Email: {*hksh,clayscot,xuanlong*}*@umich.edu*

as consistency and asymptotic normality. Our developed theory shows that ACS's algorithm has the same asymptotic rate of convergence as the MLE.

There are two common asymptotic regimes in geostatistics: *increasing-domain* and *fixed-domain*. In the former setting which is suitable for assessing the impacts of the spatial geometry of samples on estimating the covariance parameters, the smallest distance among the sampling points is bounded away from zero and more samples are collected by increasing the diameter of the spatial domain. In the latter regime, which carries more informative about interpolation procedure, the data are sampled in a fixed, bounded domain and the observations get denser as the sample size increases. This paper studies the increasing-domain asymptotics of ACS's algorithm

2. Problem formulation and ACS's algorithm

Let $\mathfrak{G} : \mathbb{R}^d \mapsto \mathbb{R}$ be a real valued, zero mean and stationary GP whose covariance function is given by

$$\mathbb{E}\mathfrak{G}(s)\,\mathfrak{G}(s') = \phi_0 K(s - s', \theta_0), \quad \forall\, s, s' \in \mathbb{R}^d. \tag{2.1}$$

The scalar $\phi_0 \in \mathcal{I}$ represents the variance of $\mathfrak{G}$ and the $m-$tuple $\theta_0 \in \Theta$ stands for the unknown correlation parameters such as the range or smoothness parameters. It is assumed throughout this paper that $\Theta \subset \mathbb{R}^m$ and $\mathcal{I} \in (0, \infty)$ are compact with respect to the Euclidean topology. Estimating (ϕ_0, θ_0) given n samples of one realization of $\mathfrak{G}$ is an objective of numerous applications in geostatistics and machine learning. Specifically, $\mathfrak{G}$ is observed at the locations $\mathcal{D}_n = \{s_1, \cdots, s_n\}$ and the collected samples form a column vector $Y = [\mathfrak{G}(s_1), \cdots, \mathfrak{G}(s_n)]^T$ of length n. We now rigorously present the geometric structure of $\mathcal{D}_n$.

Assumption 2.1. Suppose that there is $N \in \mathbb{N}$ such that $n = N^d$. There exists $\delta \in \left[0, \frac{1}{2}\right)$ for which $\mathcal{D}_n$ is a *$d-$dimensional $\delta-$perturbed regular lattice* (with unit grid size). Namely,

$$\mathcal{D}_n = \left\{ v_i + \delta p_i : \; v_i \in \mathcal{V}_{N,d},\; p_i \in [-1, 1]^d \right\}_{i=1}^{n},$$

in which $\mathcal{V}_{N,d} \coloneqq \{v_1, \cdots, v_n\} = \{1, \ldots, N\}^d$ denotes the $d-$dimensional regular lattice.

The discrepancy of $\mathcal{D}_n$ from the $d-$dimensional regular lattice is controlled by δ. $\mathcal{D}_n$ forms a regular lattice for $\delta = 0$ and the irregularity becomes more discernible as it increases. Let us introduce ACS's estimation algorithm proposed in [1].

$$\left(\hat{\phi}_n, \hat{\theta}_n\right) = \underset{(\phi,\theta)\in\mathcal{I}\times\Theta}{\arg\max}\; F_n(Y, \phi, \theta), \tag{2.2}$$

$$F_n(Y, \phi, \theta) \coloneqq \frac{1}{n}\left(\phi Y^T K_n(\theta)\, Y - \frac{\phi^2}{2} \left\| K_n(\theta) \right\|_{\ell_2}^2\right).$$

Unlike the log-likelihood loss function, $F_n(Y, \phi, \theta)$ has no dependence to the *Cholesky* factorization of the correlation matrix $K_n(\theta)$ and can be computed in $\mathcal{O}\left(n^2\right)$ operations, even for the irregularly spaced lattices. Another remarkable advantage

of using $F_n(Y,\phi,\theta)$ over the log-likelihood function is that it can be evaluated without storing the n by n correlation matrix.

3. Main results

This section is devoted to establish the asymptotic characteristics of the optimization problem (2.2). We first introduce an alternative formulation for $(\hat{\phi}_n, \hat{\theta}_n)$ giving a new perspective on the numerical difficulties of solving (2.2). The quadratic and concave form of $F_n(Y,\phi,\theta)$ in terms of ϕ yields a closed form solution for $\hat{\phi}_n$ in terms of Y and $\hat{\theta}_n$. That is,

$$\hat{\phi}_n = \left\| K_n(\hat{\theta}_n) \right\|_{\ell_2}^{-2} \left(Y^T K_n(\hat{\theta}_n) Y \right).$$

In which $\hat{\theta}_n$ is given by

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} G_n(Y,\theta), \quad \text{where} \quad G_n(Y,\theta) = \|K_n(\theta)\|_{\ell_2}^{-1} \left(Y^T K_n(\theta) Y \right). \tag{3.1}$$

As it is apparent from (3.1), $\hat{\theta}_n$ is the global maximizer of a nonconcave objective function $G_n(Y,\theta)$ even for the simplest scenarios of the isotropic Matern covariances. However the main result of this section shows that all the stationary points of $G_n(Y,\theta)$ concentrated around a small neighborhood of θ_0 with high probability. Namely, despite the non-concavity of G_n, it still retains the crucial properties of concave function. So, a highly accurate approximation of $\hat{\theta}_n$ can be obtained using the common optimization techniques.

Assumption 3.1. The following conditions are satisfied by Ω and K.

(A1) Θ and $\mathcal{I}$ are *compact connected* subsets of $\mathbb{R}^m$ and $(0,\infty)$, respectively.

(A2) There are bounded scalars $M > 0$ and $r_1 > 1$ such that for any $s \in \mathcal{D}_n$,

$$\max_{s'\in\mathcal{D}_n(s,r_1)} |K(s'-s,\theta_2) - K(s'-s,\theta_1)| \geq M \|\theta_2 - \theta_1\|_{\ell_2}, \quad \forall\, \theta_1,\theta_2 \in \Theta.$$

(A3) For some $q \in \{2,3\}$, there exists a positive scalar $C_{K,\Theta}$ such that

$$\max_{\theta\in\Theta} \left[|K(s,\theta)| \vee \left| \frac{\partial}{\partial\theta_{j_1}} \cdots \frac{\partial}{\partial\theta_{j_q}} K(s,\theta) \right| \right] \leq \frac{C_{K,\Theta}}{1 + \|s\|_{\ell_2}^{d+1}},$$
$$\forall\, s \in \mathbb{R}^m, q \quad j_1,\ldots,j_q \in \{1,\ldots,m\}.$$

Based upon Propositions 3.1 – 3.3, Assumption 3.1 holds for the popular classes of geometric anisotropic covariance functions such as Matern, powered exponential and rational quadratic.

Theorem 3.1. *Suppose that $\mathcal{D}_n$ admits Assumption 2.1 and Assumption 3.1 with $q = 2$ in (A3) holds for Ω and K. Then for some appropriately chosen constant $C > 0$, any stationary point of (2.2) satisfies*

$$\lim_{n\to\infty} \mathbb{P}\left(\left\| \hat{\theta}_n - \theta_0 \right\|_{\ell_2} \vee \left| \frac{\hat{\phi}_n}{\phi_0} - 1 \right| \geq C\sqrt{\frac{\log n}{n}} \right) = 0,$$

The other asymptotic properties of ACS's algorithm are omitted due to the space constraints. We refer the reader to [2] for a thorough analysis.

4. Numerical studies

We conduct a simulation study to assess the statistical performance of the optimization problem 2.2. For each experiment we generate a zero mean stationary GP on $\mathbb{R}^2$ with the covariance function

$$\operatorname{cov}\left(\mathfrak{G}\left(s\right),\mathfrak{G}\left(s'\right)\right)=\sigma_0^2 K\left(\left\|B_0\left(s-s'\right)\right\|_{\ell_2}\right),\quad s,s'\in\mathbb{R}^2;\quad B_0=\begin{pmatrix}\theta_0^{-1} & 0\\ 0 & \rho_0^{-1}\end{pmatrix}.$$

We choose the correlation function $K\left(\cdot\right)$ to be either rational quadratic or Matern, whose exact forms are respectively given from left to right as

$$K\left(u\right)=\left(1+u^2\right)^{-\left(1+\nu_0\right)},\quad K\left(u\right)=\frac{2^{1-\nu_0}u^{\nu_0}}{\Gamma\left(\nu_0\right)}\mathfrak{K}_{\nu_0}\left(u\right).$$

Here the smoothness parameter ν_0 is assumed to be known and $\mathfrak{K}_{\nu_0}\left(\cdot\right)$ denotes the modified Bessel function of the second kind associated with ν_0. We observe $\mathfrak{G}$ on a randomly perturbed regular lattice of size 100^2 with $\delta\in\{0.1,0.3\}$. The goal is to construct a confidence interval for the estimated parameters $(\sigma_0,\rho_0,\theta_0)$. The root mean squared error (RMSE) in Table 1 is computed using 100 independent experiments. As it is apparent from Table 1, σ_0 has a considerably narrower confidence interval than that of ρ_0 and θ_0.

	$\delta=0.1$	$\delta=0.3$
Matern covariance ($\nu_0=0.5$)	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.988\pm0.096$ $\hat{\rho}\pm\text{RSME}=6.042\pm1.885$ $\hat{\theta}\pm\text{RSME}=4.091\pm1.110$	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.993\pm0.097$ $\hat{\rho}\pm\text{RSME}=6.478\pm1.908$ $\hat{\theta}\pm\text{RSME}=4.038\pm1.272$
Matern covariance ($\nu_0=1.5$)	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.993\pm0.108$ $\hat{\rho}\pm\text{RSME}=05.965\pm1.981$ $\hat{\theta}\pm\text{RSME}=3.740\pm1.146$	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.984\pm0.104$ $\hat{\rho}\pm\text{RSME}=6.160\pm1.890$ $\hat{\theta}\pm\text{RSME}=3.970\pm1.243$
Rational quadratic covariance ($\nu_0=0.5$)	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.992\pm0.071$ $\hat{\rho}\pm\text{RSME}=5.978\pm1.241$ $\hat{\theta}\pm\text{RSME}=4.092\pm0.843$	$(\sigma_0,\rho_0,\theta_0)=(1,6'4)$ $\hat{\sigma}\pm\text{RSME}=0.989\pm0.076$ $\hat{\rho}\pm\text{RSME}=5.921\pm1.208$ $\hat{\theta}\pm\text{RSME}=4.037\pm1.064$
Rational quadratic covariance ($\nu_0=1.5$)	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.996\pm0.036$ $\hat{\rho}\pm\text{RSME}=6.116\pm0.821$ $\hat{\theta}\pm\text{RSME}=4.045\pm0.543$	$(\sigma_0,\rho_0,\theta_0)=(1,6,4)$ $\hat{\sigma}\pm\text{RSME}=0.998\pm0.036$ $\hat{\rho}\pm\text{RSME}=6.158\pm0.766$ $\hat{\theta}\pm\text{RSME}=4.150\pm0.524$

Table 1. Mean and RMSE of estimated parameters over 100 independent experiments for the geometric anisotropic covariance functions, where $\mathcal{D}_n$ is a perturbed lattice of size 100^2 with associated $\delta\in\{0.1,0.3\}$.

5. Conclusion

This paper summarizes the comprehensive increasing-domain asymptotic study of ACS's inversion-free algorithm in [2]. To our knowledge, [2] is among the first asymptotic analysis of the inversion-free optimization-based techniques for estimating the covariance parameters.

References

[1] M. Anitescu, J. Chen, and M. L. Stein, "An inversion-free estimating equation approach for Gaussian process models", *submitted for publication* (2014).

[2] H. Keshavarz, C. Scott, and X.L. Nguyen. "On the consistency of inversion-free parameter estimation for Gaussian random fields", *arXiv preprint* arXiv:1601.03822 (2016).

[3] M.L. Stein, J. Chen, and M. Anitescu, "Difference filter preconditioning for large covariance matrices", *SIAM Journal on Matrix Analysis and Applications* 33, no. 1 (2012): 52–72.

Efficient High Dimensional Interaction Search

Gian-Andrea Thanei

(joint work with Rajen Shah, Nicolai Meinshausen)

We study the interaction search problem from a computational point of view. We assume we are given p feature vectors $X_1, ..., X_p$ that have binary entries in $\{-1, 1\}$. We want to find product interactions of the form $X_l X_k$, we do so by studying the interaction frequency of each pair (l, k)

$$F_{lk} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i = X_{il} X_{ik}\}}.$$

The goal is to find the pair (l^*, k^*) for which $F_{l^* k^*}$ is maximal.

Exhaustively scanning through all pairs (l, k) has a complexity of $\mathcal{O}(np^2)$. For large p this is unfeasible.

We translate the interaction search problem to a nearest neighbor problem by writing

$$F_{lk} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i = X_{il} X_{ik}\}} = 1 - \frac{1}{4n} \sum_{i=1}^{n} (Y_i X_{il} - X_{ik})^2 = 1 - D_{lk}$$

Instead of maximizing the interaction frequency F_{lk} we minimize the distance D_{lk}. In practice this is done by building a matrix Z, as $Z_{il} = X_{il} Y_i$. The minimum distance D_{lk} is then found by looking for the closest columns among X and Z. This still has a complexity of $\mathcal{O}(np^2)$. We reduce the complexity by projecting both X and Z down to a p-dimensional vector using a random projection $R \in \mathbb{R}^n$,(for example $R \sim \mathcal{N}(0, I_{n \times n})$). We then perform nearest neighbor search on $R^T X$ and $R^T Z$, by first sorting both vectors and then stepping through them exploiting the spatial structure of sorted vectors and thereby finding the p closest pairs in the vectors $R^T X$ and $R^T Z$. This has complexity $\mathcal{O}(p \log(p))$.

To demonstrate that in fact this can offer a significant speed up we consider the case of a perfect interaction, i.e. $F_{l^*k^*} = 1$. This implies $Y_i = X_{il^*}X_{ik^*}$ $\forall i \in \{1, ..., n\}$, therefore $D_{l^*k^*} = 0$. Even the projected pair still has a distance 0:

$$|R^T Z_{\cdot l^*} - R^T X_{\cdot k^*}| = |R^T (Z_{\cdot l^*} - X_{\cdot k^*})| = 0,$$

whereas all the other pairs will be further apart. This means we can recover the closest pair using only one projection. Thus we can find the pair with the perfect interaction frequency $F_{l^*k^*} = 1$ in $\mathcal{O}(np)$ operations.

For weaker interactions one repeats the projection step multiple times, thereby boosting the probability of discovering a true interaction in the projected data.

Some recent Developments in Post-Selection inference

Robert Tibshirani

We describe the problem of "selective inference". This addresses the following challenge: having mined a set of data to find potential associations, how do we properly assess the strength of these associations? The fact that we have "cherry picked" – searched for the strongest associations – means that we must set a higher bar for declaring significant the associations that we see. This challenge becomes more important in the era of big data and complex statistical modeling: the cherry tree (dataset) can very large and the tools for cherry picking (statistical learning methods) are now very sophisticated. We describe some recent new developments in selective inference and illustrate their use in forward stepwise regression, the lasso, and principal components analysis.

This is joint work with many people including Jonathan Taylor, Richard Lockhart, Ryan Tibshirani, Will Fithian, Jason Lee, Yuekai Sun, Dennis Sun, Yun Jun Choi, Max G'Sell, Stefan Wager, and Alex Chouldechova.

Inferring High-Dimensional Poisson Autoregressive Models

Rebecca Willett
(joint work with Eric Hall, Garvesh Raskutti)

Time series count data arise in a variety of applications (cf. [1, 2, 3]), one of the most notable being biological neural networks [4, 5, 6, 7, 8]. In this application, we record the times at which each neuron in the network fires or "spikes" and wish to infer the structure of the underlying network. Action potentials or neuron spikes can trigger or inhibit spikes in connected neurons, so understanding excitation and inhibition among neurons provides key insight into the structure and operation of the neural network. A central question in the design of this experiment is *"for how long must I collect data before I can be confident that my inference of the neural network is accurate?"* Clearly the answer to this question will depend not only on the number of neurons being recorded, but also on what we may assume *a priori* about the network. Unfortunately, existing statistical and machine learning theory does not address this problem, which is the focus of this work.

This example of a biological neural network can be modeled as an auto-regressive point process. That is, at each time t, we observe a high-dimensional vector of counts, and the distribution of those counts depends on previous observations. Inferring these dependencies is a key challenge in many settings because a precise understanding of these dependencies facilitates more accurate predictions and interpretable models of the forces that determine the distribution of each new observation.

This work focuses on multivariate settings, particularly where the vector observed at each time is high-dimensional relative to the duration of the time series. We conduct a detailed investigation of a particular time series count data model: the *vector log-linear Poisson autoregressive* (PAR) model. The PAR model has been explicitly studied in [9, 10, 11], is closely related to the discrete-time INGARCH model [12, 13], and can be considered a discretized version of the continuous-time Hawkes point process model [14, 15]. We focus specifically on estimating the parameters of a vector PAR model from a time series of count data by using a regularized maximum likelihood estimation approach that generalizes past work on Poisson inverse problems (cf. [16, 17, 18]). While similar algorithms have been proposed in the above-mentioned PAR literature, little is known about their *sample complexity* or *how inference accuracy scales with the key parameters such as the size of the network, the time spent collecting observations, and the density of edges within the network or dependencies among entities.* The temporal dependence among events can make such analyses particularly challenging and beyond the scope of much current research in high-dimensional statistical inference (see [19] for an overview).

That said, there has been a large body of work providing theoretical results for certain high-dimensional models under low-dimensional structural constraints (see e.g. [18, 20, 21, 22, 23, 24, 25]). The majority of prior work has focussed on the setting where samples are independent and/or follow a Gaussian distribution; this work exploits many properties of linear systems and Gaussian random variables that can not be applied to non-Gaussian and non-linear auto-regressive models. In the Poisson auto-regressive model, we have dependent count data samples and signal-dependent Poisson noise. [20, 18, 26] provide results for non-Gaussian noise but still rely on independent observations. Another method [27] studies a general framework for point process (including the Hawkes process) and provides estimation bounds for a LASSO-type estimator. Our work emphasizes the high-dimensional setting and bounds for short-duration time series.

In this work, we develop performance guarantees for the vector PAR model that provide sample complexity guarantees in the high-dimensional setting under low-dimensional structural assumptions such as sparsity of the underlying auto-regressive parameter matrix. In particular, our main contribution is the derivation of mean-squared-error bounds on the proposed estimator as a function of the problem dimension, sparsity, and the number of observations in time.

We consider the log-linear vector Poisson autoregressive model:

$$X_{t+1}|X_t \sim \text{Poisson}(e^{\nu - A^* X_t}), \tag{1.1}$$

where $(X_t)_{t=0}^T$ are M-variate observation vectors, $A^* \in [0, A_{\max}]^{M\times M}$ is some unknown parameter matrix, and $\nu \in [\nu_{\min}, \nu_{\max}]^M$ is a known rate parameter[1]. A similar model appears in [11], but that work focuses on maximum likelihood and weighted least squares estimators in univariate settings that are known to perform poorly in high-dimensional settings (as is our focus).

Under this model, the conditional likelihood can be expressed explicitly as:

$$\mathbb{P}(X_{t+1} = y | X_t = x, A) = \prod_{m=1}^{M} \frac{\exp(-e^{\nu_m - A_m^\top x}) e^{(\nu_m - A_m^\top x) y_m}}{y_m!},$$

where x and y are M-variate vectors and $A_m^\top$ is the m^{th} row of a candidate parameter matrix A.

In general, we observe T samples $(X_t)_{t=0}^T$ and our goal is to infer the matrix A^*. In the setting where M is large, we need to impose additional assumptions on A^* in order to have strong performance guarantees. In particular we assume that the matrix A^* is s-sparse, meaning that A^* belongs to the following class:

$$\mathcal{M}_S = \{A \in [0, A_{\max}]^{M\times M} \mid \sum_{\ell=1}^{M}\sum_{m=1}^{M} \mathbf{1}(|A_{\ell,m}| \neq 0) \leq s\}.$$

Finding an optimal estimator within this parameter class would use an ℓ_0 penalty to ensure that the estimator has at most s non-zero entries. However, this is a difficult optimization problem due to the non-convexity of the ℓ_0 function. Therefore, we instead find an estimator using the element-wise ℓ_1 decomposable regularizer, the convex relaxation of the ℓ_0 function, along with the negative log likelihood (using the known, constant vector ν):

$$\widehat{A} = \arg \min_{A\in[0,A_{\max}]^{M\times M}} \frac{1}{T}\sum_{t=0}^{T-1}\sum_{m=1}^{M}\left(e^{\nu_m - A_m^\top X_t} + A_m^\top X_t X_{m,t+1}\right) + \lambda\|A\|_{1,1} \tag{1.2}$$

where $\|\cdot\|_{1,1}$ is the element-wise ℓ_1 norm:

$$\|A\|_{1,1} = \sum_{\ell=1}^{M}\sum_{m=1}^{M} |A_{\ell,m}|.$$

The above is the regularized maximum likelihood estimator (RMLE), which attempts to find an estimate of A which both fits the data and has many zero valued elements. The goal is to derive bounds for $\|\widehat{A} - A^*\|_F^2$, the difference between the regularized maximum likelihood estimator, $\widehat{A}$, and the true generating network, A^*, under the assumption that the true network is sparse, which is our main theorem. As we will see in the main theorem, a crucial quantity in the regret bounds is the value ρ which is the maximum number of non-zeros in any row of A^*. If ρ is considered a constant independent of M and s, then good error bounds can be determined, but if ρ is on the order of s, meaning there's a single, dense row of non-zeros, poor error rates will be observed.

[1]Our framework easily extends to unknown ν but the notation is cumbersome; we focus on known ν for simplicity of presentation.

Theorem: *Let* $\widehat{A}$ *be the RMLE as in Equation 1.2, and assume* $\lambda \geq ||\frac{2}{T}\sum_{t=0}^{T-1}(X_{t+1} - e^{A^*X_t})X_t^\top\|_{\infty,\infty}$ *where* $\|.\|_{\infty,\infty}$ *is the element-wise* ∞*-norm then,*

$$\|\widehat{A} - A^*\|_F^2 \leq O(e^\rho s\lambda^2)$$

with probability at least $1 - 2\exp(-\min\left(c_3T/\rho^2 - c_4 s\log(2M), c_2MT\right))$ *where* c_2, c_3 *and* c_4 *are independent of* M, T, ρ *and* s*. Further,*

$$\|\frac{2}{T}\sum_{t=0}^{T-1}(X_{t+1} - e^{A^*X_t})X_t^\top\|_\infty \leq \frac{8C_1^2e^{\nu_{\max}}\log^3(MT)}{\sqrt{T}}$$

with high probability yielding the overall error rate of

$$\|\hat{A} - A^*\|_F^2 \leq O\left(\frac{e^\rho s}{T}\log^6(MT)\right)$$

with probability at least $1 - \exp(-c_6\min\left(T/\rho^2 - s\log(M), \log(MT)\right))$ *for some* c_6 *independent of* M, T, ρ *and* s*.*

References

[1] Kurt Brännäs and Per Johansson. Time series count data regression. *Communications in Statistics-Theory and Methods*, 23(10):2907–2925, 1994.

[2] Scott L Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

[3] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013.

[4] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, 2004.

[5] M. Hinne, T. Heskes, and M. A. J. van Gerven. Bayesian inference of whole-brain networks. *arXiv:1202.1696 [q-bio.NC]*, 2012.

[6] M. Ding, CE Schroeder, and X. Wen. Analyzing coherent brain networks with Granger causality. In *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pages 5916–8, 2011.

[7] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454:995–999, 2008.

[8] M. S. Masud and R. Borisyuk. Statistical technique for analysing functional connectivity of multiple spike trains. *Journal of Neuroscience Methods*, 196(1):201–219, 2011.

[9] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

[10] Fukang Zhu and Dehui Wang. Estimation and testing for a poisson autoregressive model. *Metrika*, 73(2):211–230, 2011.

[11] Konstantinos Fokianos and Dag Tjøstheim. Log-linear poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578, 2011.

[12] Andréas Heinen. Modelling time series count data: an autoregressive conditional poisson model. *Available at SSRN 1117187*, 2003.

[13] Fukang Zhu. Modeling overdispersed or underdispersed count data with generalized poisson integer-valued garch models. *Journal of Mathematical Analysis and Applications*, 389(1):58–71, 2012.

[14] A. G. Hawkes. Point spectra of some self-exciting and mutually-exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:83–90, 1971.

[15] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes, Vol. I: Probability and its Applications*. Springer-Verlag, New York, second edition, 2003.

[16] M. Raginsky, R. Willett, Z. Harmany, and R. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010. arXiv:0910.5146.
[17] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank. Performance bounds for expander-based compressed sensing in Poisson noise. *IEEE Transactions on Signal Processing*, 59(9), 2011. arXiv:1007.2377.
[18] X. Jiang, R. Willett, and G. Raskutti. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 61:4458–4474, 2015.
[19] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
[20] S. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36:614–636, 2008.
[21] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.
[22] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2010.
[23] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:6976–6994, 2011.
[24] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13:398–427, 2012.
[25] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567, 2015.
[26] X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. Willett. A data-dependent weighted lasso under poisson noise. *arXiv preprint arXiv:1509.08892*, 2015.
[27] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 02 2015.

Estimating whole brain dynamics using spectral clustering

Yi Yu
(joint work with Ivor Cribben)

The estimation of time-varying networks for functional Magnetic Resonance Imaging (fMRI) data sets is of increasing importance and interest. In this work, we formulate the problem in a high-dimensional time series framework, and introduce a data-driven method, namely Network Change Points Detection (NCPD), which detects change points in the network structure of a multivariate time series, with each component of the time series represented by a node in the network. NCPD is applied to various simulated data and a resting-state fMRI data set. The new methodology also allows us to identify common functional states within and across subjects. Finally, NCPD promises to offer a deep insight into the large-scale characterisations and dynamics of the brain.

Statistics in Big Data Optimization

Tong Zhang
(joint work with Shai Shalev-Schwartz, Rie Johnson, et al)

Due to the explosion of data in our society, it is necessary to design computational algorithms that can solve big data optimization problems. One example is the click through rate (CTR) prediction problem in computational advertizing, which is at the heart of efficient automatic advertizing systems used in the modern internet industry. When a user views some online content, internet companies will display ads to match the user's need, so that the user's chance of clicking the displayed ads will be maximized. The CTR estimation problem can be solved using the following linear logistic regression (or its nonlinear variant):

$$\min_w \frac{1}{n}\sum_{i=1}^{n}\left[\ln(1+e^{-w^\top x_i y_i})\right] + \frac{\lambda}{2}\|w\|_2^2,$$

where the data are represented as (x_i, y_i) with $y_i \in \{\pm 1\}$ indicating whether a user clicks on an ad, and w is a model parameter.

Due to the automated data gathering mechanism on the internet, the training data for this optimization problem can be as many as 100 billion, and the dimensionality can also reach as high as 100 billion. These modern big-data machine learning problems encountered in the industry involve optimization problems so large that traditional methods are difficult to handle. More generally, we are interested in solving the following abstract problem:

$$\min_w f(w), \qquad f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w),$$

where or logistic problem, $f_i(w) = \ln(1+e^{-w^\top x_i y_i})$. Here big data means that n can be extremely large.

The complex issues in solving these large scale applications have stimulated fast development of novel optimization techniques in recent years. In particular, I will demonstrate how novel applications of statistical thinking can be used to efficiently solve these large scale optimization problems. In particular, I discuss several statistical ideas to alleviate the computational challenge via sampling methods that can be used to select a small number of the most important data points for the problem, including modern stochastic optimization such as those in [1, 2], and some other statistical techniques that are used in practice, but not published.

References

[1] Shai Shalev-Shwartz and Tong Zhang, *Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization*, JMLR 2013.

[2] Rie Johnson and Tong Zhang, *Accelerating Stochastic Gradient Descent using Predictive Variance Reduction*, NIPS 2013.

Reporter: Gilles Blanchard

Participants

Prof. Dr. Francis Bach
INRIA / ENS
CS 81321
23, Avenue d'Italie
75214 Paris Cedex 13
FRANCE

Dr. Krishnakumar Balasubramanian
Department of Statistics
University of Wisconsin
MSC, 1300 University Ave.
Madison WI 53706-1685
UNITED STATES

Dr. Quentin Berthet
Department of Pure Mathematics
and Mathematical Statistics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Gilles Blanchard
Fachbereich Mathematik
Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
GERMANY

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

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

Alexandra Carpentier
Department of Pure Mathematics
and Mathematical Statistics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Rui M. Castro
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
NETHERLANDS

Prof. Dr. Mathias Drton
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
UNITED STATES

Prof. Dr. John Duchi
Statistics and Electrical Engineering
Stanford University
Sequoia Hall
Stanford CA 94305
UNITED STATES

Prof. Dr. Rina Foygel Barber
Department of Statistics
The University of Chicago
5734 University Avenue
Chicago, IL 60637-1514
UNITED STATES

Prof. Dr. Elisabeth Gassiat
Laboratoire de Mathématiques
Université Paris Sud (Paris XI)
Batiment 425
91405 Orsay Cedex
FRANCE

Prof. Dr. Christophe Giraud
Centre de Mathématiques Appliquées
UMR 7641 - CNRS
École Polytechnique
91128 Palaiseau Cedex
FRANCE

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

Christina Heinze
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Martin Jaggi
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. John D. Lafferty
Department of Statistics and Computer Science
The University of Chicago
Jones 120 B
5747 S. Ellis Avenue
Chicago, IL 60637
UNITED STATES

Prof. Dr. Jing Lei
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

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

Dr. Po-Ling Loh
Department of Statistics
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104
UNITED STATES

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

Prof. Dr. Andrea Montanari
Department of Electrical Engineering
Stanford University
Stanford, CA 94305-4055
UNITED STATES

Dr. Sach Mukherjee
Deutsches Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE)
Ernst-Robert-Curtius-Straße 12
53117 Bonn
GERMANY

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

Prof. Dr. Sahand Negahban
Department of Statistics
Yale University
Yale Station
P.O. Box 2179
New Haven, CT 06520-2179
UNITED STATES

Prof. Dr. Robert D. Nowak
Department of Electrical & Computer Engineering
University of Wisconsin-Madison
Engineering Hall # 3627
1415 Engineering Drive
Madison, WI 53706
UNITED STATES

Dr. Guillaume Obozinski
Ecole des Ponts - ParisTech
Equipe Imagine
Champs-sur-Marne, Cité Descartes
6, av. Blaise Pascal
77455 Marne-le-Vallée Cedex 2
FRANCE

Emilija Perkovic
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Jonas Peters
Max Planck Institute for Intelligent Systems
Spemannstraße 41
72076 Tübingen
GERMANY

Garvesh Raskutti
Department of Statistics
University of Wisconsin, MSC
1300 University Avenue
Madison, WI 53706-1685
UNITED STATES

Prof. Dr. Peter Richtarik
School of Mathematics
University of Edinburgh
James Clerk Maxwell Bldg.
Edinburgh EH9 3JZ
UNITED KINGDOM

Prof. Dr. Philippe Rigollet
Department of Mathematics
Center for Statistics
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
Lehrstuhl für Stochastik
Ruhr-Universität Bochum
44780 Bochum
GERMANY

Prof. Dr. Lorenzo Rosasco
Massachusetts Institute of Technology
Office: 46-5155 C
43 Vassar Street
Cambridge, MA 02139
UNITED STATES

Prof. Dr. Saharon Rosset
Department of Mathematics
School of Mathematical Sciences
Tel Aviv University
P.O.Box 39040
Ramat Aviv, Tel Aviv 69978
ISRAEL

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

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

Prof. Dr. Clayton Scott
Electrical Engineering & Computer Science Dept.
The University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122
UNITED STATES

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

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. Robert Tibshirani
Department of Statistics
Stanford University
Sequoia Hall
Stanford, CA 94305-4065
UNITED STATES

Prof. Dr. Ryan Tibshirani
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

Tengyao Wang
Statistics Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

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

Prof. Dr. Rebecca Willett
Electrical and Computer Engineering Dept.
University of Wisconsin-Madison
3627 Engineering Hall, Rm. 3537
1415 Engineering Drive
Madison, WI 53706
UNITED STATES

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

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

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

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

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