THE ANNALS

of

STATISTICS

AN OFFICIAL JOURNAL OF THE

INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Convergence of de Finetti’s mixing measure in latent structure models for observed exchangeable sequences ......... YUN WEI AND XUANLONG NGUYEN 1859
Detecting multiple replicating signals using adaptive filtering procedures JINGSHU WANG, LIN GUI, WEIJIE J. SU, CHIARA ŠABATTI AND ART B. OWEN 1890
Optimal signal detection in some spiked random matrix models: Likelihood ratio tests and linear spectral statistics .......... DEBAPRATIM BANERJEE AND ZONGMING MA 1910
On universally consistent and fully distribution-free rank tests of vector independence HONGJIAN SHI, MARC HALLIN, MATHIAS DRTON AND FANG HAN 1933
Limit theorems for distributions invariant under groups of transformations MORGANE AUSTERN AND PETER ORBANZ 1960
Distributed adaptive Gaussian mean estimation with unknown variance: Interactive protocol helps adaptation ................. T. TONY CAI AND HONGJI WEI 1992
Asymptotic accuracy of the saddlepoint approximation for maximum likelihood estimation ........................................ JESSE GOODMAN 2021
Generalization error bounds of dynamic treatment regimes in penalized regression-based learning ................. EUN JEONG OH, MIN QIAN AND YING KUEN CHEUNG 2047
A CLT for second difference estimators with an application to volatility and intensity EMIL A. STOLTENBERG, PER A. MYKLAND AND LAN ZHANG 2072
Sharp optimal recovery in the two component Gaussian mixture model MOHAMED NDAOUD 2096
Precise statistical analysis of classification accuracies for adversarial training ADEL JAVANMARD AND MAHDI SOLTANOLKOTABI 2127
Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices ........................................ JAOUD Mourtada 2157
Stochastic continuum-armed bandits with additive models: Minimax regrets and adaptive algorithm ......................... T. TONY CAI AND HONGMING PU 2179
Asymptotic independence of spiked eigenvalues and linear spectral statistics for large sample covariance matrices .......... ZHIXIANG ZHANG, SHURONG ZHENG, GUANGMING PAN AND PING-SHOU ZHONG 2205
Learning mixtures of permutations: Groups of pairwise comparisons and combinatorial method of moments ............... CHENG MAO AND YIHONG WU 2231
Generalized resilience and robust statistics BANGHUA ZHU, JIANTAO JIAO AND JACOB STEINHARDT 2256
Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses .................. HUIJIE FENG, YANG NING AND JIWEI ZHAO 2284
On the robustness of minimum norm interpolators and regularized empirical risk minimizers . GEOFFREY CHINOT, MATTHIAS LOFFLER AND SARA VAN DE GEER 2306
Random graph asymptotics for treatment effect estimation under network interference SHUANGNING LI AND STEFAN WAGER 2334
An \(\ell_p\) theory of PCA and spectral clustering EMMANUEL ABBE, JIANQING FAN AND KAIZHENG WANG 2359
Estimation of smooth functionals in high-dimensional models: Bootstrap chains and Gaussian approximation .............................. VLADIMIR KOLTCHINSKII 2386
A minimax framework for quantifying risk-fairness trade-off in regression
Evgenii Chzhen and Nicolas Schreuder 2416
Consistency of invariance-based randomization tests ............... Edgar Dobriban 2443

Erratum
Asymptotic genealogies of interacting particle systems with an application to sequential
Monte Carlo ............................ Jere Koskela, Paul A. Jenkins,
Adam M. Johansen and Dario Spanò 2467
CONVERGENCE OF DE FINETTI’S MIXING MEASURE IN LATENT STRUCTURE MODELS FOR OBSERVED EXCHANGEABLE SEQUENCES

BY YUN WEI¹,a AND XUANLONG NGUYEN²,b

¹Department of Mathematics, University of Michigan, acloudwei@umich.edu
²Department of Statistics, University of Michigan, bxuanlong@umich.edu

Mixtures of product distributions are a powerful device for learning about heterogeneity within data populations. In this class of latent structure models, de Finetti’s mixing measure plays the central role for describing the uncertainty about the latent parameters representing heterogeneity. In this paper, posterior contraction theorems for de Finetti’s mixing measure arising from finite mixtures of product distributions will be established; under the setting the number of exchangeable sequences of observed variables increases while sequence length(s) may be either fixed or varied. The role of both the number of sequences and the sequence lengths will be carefully examined. In order to obtain concrete rates of convergence, a first-order identifiability theory for finite mixture models and a family of sharp inverse bounds for mixtures of product distributions will be developed via a harmonic analysis of such latent structure models. This theory is applicable to broad classes of probability kernels composing the mixture model of product distributions for both continuous and discrete domain X. Examples of interest include the case the probability kernel is only weakly identifiable in the sense of (Ann. Statist. 44 (2016) 2726–2755), the case where the kernel is itself a mixture distribution as in hierarchical models, and the case the kernel may not have a density with respect to a dominating measure on an abstract domain X, such as Dirichlet processes.

REFERENCES


Key words and phrases. Mixture of product distributions, hierarchical models, inverse bounds, Fourier analysis, mixtures of repeated measurements, mixtures of grouped observations.


DETECTING MULTIPLE REPLICATING SIGNALS USING ADAPTIVE FILTERING PROCEDURES

BY JINGSHU WANG¹,a, LIN GUI¹,b, WEIJIE J. SU²,c, CHIARA SABATTI³,d AND ART B. OWEN³,e

¹Department of Statistics, The University of Chicago, a) jingshuw@uchicago.edu, b) glin6@uchicago.edu
²Department of Statistics and Data Science, University of Pennsylvania, c) suw@wharton.upenn.edu
³Department of Statistics, Stanford University, d) sabatti@stanford.edu, e) owen@stanford.edu

Replicability is a fundamental quality of scientific discoveries: we are interested in those signals that are detectable in different laboratories, different populations, across time etc. Unlike meta-analysis which accounts for experimental variability but does not guarantee replicability, testing a partial conjunction (PC) null aims specifically to identify the signals that are discovered in multiple studies. In many contemporary applications, for example, comparing multiple high-throughput genetic experiments, a large number $M$ of PC nulls need to be tested simultaneously, calling for a multiple comparisons correction. However, standard multiple testing adjustments on the $M$ PC $p$-values can be severely conservative, especially when $M$ is large and the signals are sparse. We introduce AdaFilter, a new multiple testing procedure that increases power by adaptively filtering out unlikely candidates of PC nulls. We prove that AdaFilter can control FWER and FDR as long as data across studies are independent, and has much higher power than other existing methods. We illustrate the application of AdaFilter with three examples: microarray studies of Duchenne muscular dystrophy, single-cell RNA sequencing of T cells in lung cancer tumors and GWAS for metabolomics.

REFERENCES


Key words and phrases. Simultaneous signals, meta-analysis, high-throughput experiments, composite null, multiple hypotheses testing.


OPTIMAL SIGNAL DETECTION IN SOME SPIKED RANDOM MATRIX MODELS: LIKELIHOOD RATIO TESTS AND LINEAR SPECTRAL STATISTICS

BY DEBAPRATIM BANERJEE and ZONGMING MA

Department of Statistics, The Wharton School, University of Pennsylvania, 
daban@wharton.upenn.edu,
zongming@wharton.upenn.edu

We study signal detection by likelihood ratio tests in a number of spiked random matrix models, including but not limited to Gaussian mixtures and spiked Wishart covariance matrices. We work directly with multi-spiked cases in these models and with flexible priors on signal components that allow dependence across spikes. We derive asymptotic normality for the log-likelihood ratios when the signal-to-noise ratios are below certain bounds. In addition, the log-likelihood ratios can be asymptotically decomposed as weighted sums of a collection of statistics which we call bipartite signed cycles. Based on this decomposition, we show that below the bounds we could always achieve the asymptotically optimal powers of likelihood ratio tests via tests based on linear spectral statistics which have polynomial time complexity.

REFERENCES


MSC2020 subject classifications. Primary 62C05, 62F05; secondary 60F05.

Key words and phrases. Contiguity, finite rank deformation, principal component analysis, random graphs, signal detection.


ON UNIVERSALLY CONSISTENT AND FULLY DISTRIBUTION-FREE RANK TESTS OF VECTOR INDEPENDENCE

BY HONGJIAN SHI¹,a, MARC HALLIN²,c, MATTHIAS DR r TO N¹,b AND FANG HAN³,d

¹Department of Mathematics, Technical University of Munich, a hongjian.shi@tum.de, b mathias.drton@tum.de
²ECARES and Department of Mathematics, Université Libre de Bruxelles, c mhallin@ulb.ac.be
³Department of Statistics, University of Washington, d fanghan@uw.edu

Rank correlations have found many innovative applications in the last decade. In particular, suitable rank correlations have been used for consistent tests of independence between pairs of random variables. Using ranks is especially appealing for continuous data as tests become distribution-free. However, the traditional concept of ranks relies on ordering data and is, thus, tied to univariate observations. As a result, it has long remained unclear how one may construct distribution-free yet consistent tests of independence between random vectors. This is the problem addressed in this paper, in which we lay out a general framework for designing dependence measures that give tests of multivariate independence that are not only consistent and distribution-free but which we also prove to be statistically efficient. Our framework leverages the recently introduced concept of center-outward ranks and signs, a multivariate generalization of traditional ranks, and adopts a common standard form for dependence measures that encompasses many popular examples. In a unified study, we derive a general asymptotic representation of center-outward rank-based test statistics under independence, extending to the multivariate setting the classical Hájek asymptotic representation results. This representation permits direct calculation of limiting null distributions and facilitates a local power analysis that provides strong support for the center-outward approach by establishing, for the first time, the nontrivial power of center-outward rank-based tests over root-\(n\) neighborhoods within the class of quadratic mean differentiable alternatives.

REFERENCES


LIMIT THEOREMS FOR DISTRIBUTIONS INVARIANT UNDER GROUPS OF TRANSFORMATIONS

BY MORGANE AUSTERN¹,a AND PETER ORBANZ²,b

¹Department of Statistics, Harvard University, *maustern@fas.harvard.edu
²Gatsby Computational Neuroscience Unit, University College London, *porbanz@gatsby.ucl.ac.uk

A distributional symmetry is invariance of a distribution under a group of transformations. Exchangeability and stationarity are examples. We explain that a result of ergodic theory implies a law of large numbers for such invariant distributions: If the group satisfies suitable conditions, expectations can be estimated by averaging over subsets of transformations, and these estimators are strongly consistent. We show that, if a mixing condition holds, the averages also satisfy a central limit theorem, a Berry–Esseen bound, and concentration. These are extended further to apply to triangular arrays, to randomly subsampled averages, and to a generalization of U-statistics. As applications, we obtain a general limit theorem for exchangeable random structures, and new results on stationary random fields, network models, and a class of marked point processes. We also establish asymptotic normality of the empirical entropy for a large class of processes. Some known results are recovered as special cases, and can hence be interpreted as an outcome of symmetry. The proofs adapt Stein’s method.

REFERENCES


MSC2020 subject classifications. Primary 62G20; secondary 37A30, 62M99, 60F05, 60G09.

Key words and phrases. Asymptotic normality, central limit theorems, Berry–Esseen theorems, pointwise ergodic theorems, Stein’s method, symmetry, exchangeability, ergodicity.


DISTRIBUTED ADAPTIVE GAUSSIAN MEAN ESTIMATION WITH UNKNOWN VARIANCE: INTERACTIVE PROTOCOL HELPS ADAPTATION

BY T. TONY CAI\textsuperscript{a} AND HONGJI WEI\textsuperscript{b}

\textsuperscript{a}Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, \texttt{tcai@wharton.upenn.edu}, \textsuperscript{b}\texttt{hongjiw@wharton.upenn.edu}

Distributed estimation of a Gaussian mean with unknown variance under communication constraints is studied. Necessary and sufficient communication costs under different types of distributed protocols are derived for any estimator that is adaptively rate-optimal over a range of possible values for the variance. Communication-efficient and statistically optimal procedures are developed.

The analysis reveals an interesting and important distinction among different types of distributed protocols: compared to the independent protocols, interactive protocols such as the sequential and blackboard protocols require less communication costs for rate-optimal adaptive Gaussian mean estimation. The lower bound techniques developed in the present paper are novel and can be of independent interest.

REFERENCES


Key words and phrases. Adaptive estimation, communication constraints, distributed learning, Gaussian mean, minimax lower bound, optimal rate of convergence.


ASYMPTOTIC ACCURACY OF THE SADDLEPOINT APPROXIMATION FOR MAXIMUM LIKELIHOOD ESTIMATION

BY JESSE GOODMAN

Department of Statistics, University of Auckland, jesse.goodman@auckland.ac.nz

The saddlepoint approximation gives an approximation to the density of a random variable in terms of its moment generating function. When the underlying random variable is itself the sum of \( n \) unobserved i.i.d. terms, the basic classical result is that the relative error in the density is of order \( 1/n \). If instead the approximation is interpreted as a likelihood and maximised as a function of model parameters, the result is an approximation to the maximum likelihood estimate (MLE) that can be much faster to compute than the true MLE. This paper proves the analogous basic result for the approximation error between the saddlepoint MLE and the true MLE: subject to certain explicit identifiability conditions, the error has asymptotic size \( O(1/n^2) \) for some parameters and \( O(1/n^{3/2}) \) or \( O(1/n) \) for others. In all three cases, the approximation errors are asymptotically negligible compared to the inferential uncertainty.

The proof is based on a factorisation of the saddlepoint likelihood into an exact and approximate term, along with an analysis of the approximation error in the gradient of the log-likelihood. This factorisation also gives insight into alternatives to the saddlepoint approximation, including a new and simpler saddlepoint approximation, for which we derive analogous error bounds. As a corollary of our results, we also obtain the asymptotic size of the MLE approximation error when the saddlepoint approximation is replaced by the normal approximation.

REFERENCES


MSC2020 subject classifications. Primary 62F12; secondary 41A60, 62F15.

Key words and phrases. Saddlepoint approximation, maximum likelihood estimation, approximation error.


A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage of intervention, that maps up-to-date patient information to a recommended treatment. Discovering an appropriate DTR for a given disease is a challenging issue especially when a large set of prognostic variables are observed. To address this problem, we propose penalized regression-based learning methods with $l_1$ penalty to estimate the optimal DTR that would maximize the expected outcome if implemented. We also provide generalization error bounds of the estimated DTR in the setting of finite number of stages with multiple treatment options. We first examine the relationship between value and Q-functions and derive a finite sample upper bound on the difference in values between the optimal and the estimated DTRs. For practical implementation, we develop an algorithm with partial regularization via orthogonality to construct the optimal DTR. The advantages of the proposed methods are demonstrated with extensive simulation studies and data analysis of depression clinical trials.

REFERENCES


Key words and phrases. Personalized medicine, regression-based learning, treatment decision making, variable selection, backward induction.


A CLT FOR SECOND DIFFERENCE ESTIMATORS WITH AN APPLICATION TO VOLATILITY AND INTENSITY

BY EMIL A. STOLTENBERG1,a, PER A. MYKLAND2,b AND LAN ZHANG3,c

1Department of Data Science and Analytics, BI Norwegian Business School, aemilstoltenberg@gmail.com
2Department of Statistics, The University of Chicago, bmykland@pascal.uchicago.edu
3Department of Finance, University of Illinois at Chicago, clanzhang@uic.edu

In this paper, we introduce a general method for estimating the quadratic covariation of one or more spot parameter processes associated with continuous time semimartingales, and present a central limit theorem that has this class of estimators as one of its applications. The class of estimators we introduce, that we call Two-Scales Quadratic Covariation (TSQC) estimators, is based on sums of increments of second differences of the observed processes, and the intervals over which the differences are computed are rolling and overlapping. This latter feature lets us take full advantage of the data, and, by sufficiency considerations, ought to outperform estimators that are based on only one partition of the observational window. Moreover, a two-scales approach is employed to deal with asymptotic bias terms in a systematic manner, thus automatically giving consistent estimators without having to work out the form of the bias term on a case-to-case basis. We highlight the versatility of our central limit theorem by applying it to a novel leverage effect estimator that does not belong to the class of TSQC estimators. The principal empirical motivation for the present study is that the discrete times at which a continuous time semimartingale is observed might depend on features of the observable process other than its level, such as its spot-volatility process. As an application of the TSQC estimators, we therefore show how it may be used to estimate the quadratic covariation between the spot-volatility process and the intensity process of the observation times, when both of these are taken to be semimartingales. The finite sample properties of this estimator are studied by way of a simulation experiment, and we also apply this estimator in an empirical analysis of the Apple stock. Our analysis of the Apple stock indicates a rather strong correlation between the spot volatility process of the log-prices process and the times at which this stock is traded and hence observed.

REFERENCES


MSC2020 subject classifications. Primary 60F05, 60G44, 62M09, 62M10; secondary 60G42, 62G20, 62P20.

Key words and phrases. Asynchronous times, central limit theorem, consistency, convergence rates, counting processes, endogenous observation times, high frequency, intensity, irregular times, microstructure, observed asymptotic variance, overlapping intervals, rolling intervals, sufficiency, two-scales estimation.
SHARP OPTIMAL RECOVERY IN THE TWO COMPONENT GAUSSIAN MIXTURE MODEL

By Mohamed Ndaoud

Department of Decisions Sciences, ESSEC Business School, ndaoud@essec.edu

In this paper, we study the problem of clustering in the Two component Gaussian mixture model when the centers are separated by some $\Delta > 0$. We present a nonasymptotic lower bound for the corresponding minimax Hamming risk improving on existing results. We also propose an optimal, efficient and adaptive procedure that is minimax rate optimal. The rate optimality is moreover sharp in the asymptotics when the sample size goes to infinity. Our procedure is based on a variant of Lloyd’s iterations initialized by a spectral method. As a consequence of nonasymptotic results, we find a sharp phase transition for the problem of exact recovery in the Gaussian mixture model. We prove that the phase transition occurs around the critical threshold $\tilde{\Delta}$ given by

$$\tilde{\Delta}^2 = \sigma^2 \left( 1 + \sqrt{1 + \frac{2p}{n \log n}} \right) \log n.$$

REFERENCES


MSC2020 subject classifications. Primary 62H30; secondary 62C20, 62F07.

Key words and phrases. Gaussian mixtures, sharp recovery, Lloyd’s algorithm.
PRECISE STATISTICAL ANALYSIS OF CLASSIFICATION ACCURACIES FOR ADVERSARIAL TRAINING

BY ADEL JAVANMARD\textsuperscript{1,a} AND MAHDI SOLTANOLKOTABI\textsuperscript{2,b}

\textsuperscript{1}Department of Data Sciences and Operations, University of Southern California, \textsuperscript{a}ajavanma@usc.edu
\textsuperscript{2}Department of Electrical and Computer Engineering, University of Southern California, \textsuperscript{b}soltanol@usc.edu

Despite the wide empirical success of modern machine learning algorithms and models in a multitude of applications, they are known to be highly susceptible to seemingly small indiscernible perturbations to the input data known as adversarial attacks. A variety of recent adversarial training procedures have been proposed to remedy this issue. Despite the success of such procedures at increasing accuracy on adversarially perturbed inputs or robust accuracy, these techniques often reduce accuracy on natural unperturbed inputs or standard accuracy. Complicating matters further, the effect and trend of adversarial training procedures on standard and robust accuracy is rather counter intuitive and radically dependent on a variety of factors including the perceived form of the perturbation during training, size/quality of data, model overparameterization, etc. In this paper, we focus on binary classification problems where the data is generated according to the mixture of two Gaussians with general anisotropic covariance matrices and derive a precise characterization of the standard and robust accuracy for a class of minimax adversarially trained models. We consider a general norm-based adversarial model, where the adversary can add perturbations of bounded $\ell_p$ norm to each input data, for an arbitrary $p \geq 1$. Our comprehensive analysis allows us to theoretically explain several intriguing empirical phenomena and provide a precise understanding of the role of different problem parameters on standard and robust accuracies.

REFERENCES


\textbf{MSC2020 subject classifications.} Primary 62F12, 62E20; secondary 62J12.

\textbf{Key words and phrases.} Precise high-dimensional asymptotics, adversarial training, binary classification.


EXACT MINIMAX RISK FOR LINEAR LEAST SQUARES, AND THE LOWER TAIL OF SAMPLE COVARIANCE MATRICES

BY JAOUAD MOURTADAA

CREST, ENSAE, Institut Polytechnique de Paris. ajaouad.mourtada@ensae.fr

We consider random-design linear prediction and related questions on the lower tail of random matrices. It is known that, under boundedness constraints, the minimax risk is of order $d/n$ in dimension $d$ with $n$ samples. Here, we study the minimax expected excess risk over the full linear class, depending on the distribution of covariates. First, the least squares estimator is exactly minimax optimal in the well-specified case, for every distribution of covariates. We express the minimax risk in terms of the distribution of statistical leverage scores of individual samples, and deduce a minimax lower bound of $d/(n - d + 1)$ for any covariate distribution, nearly matching the risk for Gaussian design. We then obtain sharp nonasymptotic upper bounds for covariates that satisfy a "small ball"-type regularity condition in both well-specified and misspecified cases.

Our main technical contribution is the study of the lower tail of the smallest singular value of empirical covariance matrices at small values. We establish a lower bound on this lower tail, valid for any distribution in dimension $d \geq 2$, together with a matching upper bound under a necessary regularity condition. Our proof relies on the PAC-Bayes technique for controlling empirical processes, and extends an analysis of Oliveira devoted to a different part of the lower tail.

REFERENCES


MSC2020 subject classifications. Primary 62J05; secondary 60B20, 62C20.

Key words and phrases. Least squares, covariance matrices, lower bounds, decision theory, statistical learning theory, anticoncentration.


STOCHASTIC CONTINUUM-ARMED BANDITS WITH ADDITIVE MODELS: MINIMAX REGRETS AND ADAPTIVE ALGORITHM

BY T. TONY CAIa AND HONGMING PUb

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, a tcai@wharton.upenn.edu, b hpu@wharton.upenn.edu

We consider $d$-dimensional stochastic continuum-armed bandits with the expected reward function being additive $\beta$-Hölder with sparsity $s$ for $0 < \beta < \infty$ and $1 \leq s \leq d$. The rate of convergence $\tilde{O}(s \cdot T^{\frac{\beta+1}{2\beta+1}})$ for the minimax regret is established where $T$ is the number of rounds. In particular, the minimax regret does not depend on $d$ and is linear in $s$. A novel algorithm is proposed and is shown to be rate-optimal, up to a logarithmic factor of $T$.

The problem of adaptivity is also studied. A lower bound on the cost of adaptation to the smoothness is obtained and the result implies that adaptation for free is impossible in general without further structural assumptions. We then consider adaptive additive SCAB under an additional self-similarity assumption. An adaptive procedure is constructed and is shown to simultaneously achieve the minimax regret for a range of smoothness levels.

REFERENCES


MSC2020 subject classifications. Primary 62G08; secondary 62L12.

Key words and phrases. Adaptivity, communication constraints, additive model, bandits, curse of dimensionality, minimax lower bound, optimal rate of convergence, regret, self-similarity.
ASYMPTOTIC INDEPENDENCE OF SPIKED EIGENVALUES AND LINEAR SPECTRAL STATISTICS FOR LARGE SAMPLE COVARIANCE MATRICES

BY ZHIXIANG ZHANG¹,ᵃ, SHURONG ZHENG²,ᶜ, GUANGMING PAN¹,ᵇ AND PING-SHOU ZHONG³,ᵈ

¹School of Physical and Mathematical Sciences, Nanyang Technological University, ᵃS160011@e.ntu.edu.sg, ᵇGMPAN@ntu.edu.sg
²School of Mathematics and Statistics, Northeast Normal University, ᶜzhengsr@nenu.edu.cn
³Department of Mathematics, Statistics, and Computer Science, University of Illinois Chicago, ᵈpszhong@uic.edu

We consider general high-dimensional spiked sample covariance models and show that their leading sample spiked eigenvalues and their linear spectral statistics are asymptotically independent when the sample size and dimension are proportional to each other. As a byproduct, we also establish the central limit theorem of the leading sample spiked eigenvalues by removing the block diagonal assumption on the population covariance matrix, which is commonly needed in the literature. Moreover, we propose consistent estimators of the $L_4$ norm of the spiked population eigenvectors. Based on these results, we develop a new statistic to test the equality of two spiked population covariance matrices. Numerical studies show that the new test procedure is more powerful than some existing methods.

REFERENCES


Key words and phrases. Leading spiked eigenvalues, sample covariance matrix, linear spectral statistics, central limit theorem.


LEARNING MIXTURES OF PERMUTATIONS: GROUPS OF PAIRWISE COMPARISONS AND COMBINATORIAL METHOD OF MOMENTS

BY CHENG MAO¹,a AND YIHONG WU²,b

¹School of Mathematics, Georgia Institute of Technology, ²Department of Statistics and Data Science, Yale University

In applications such as rank aggregation, mixture models for permutations are frequently used when the population exhibits heterogeneity. In this work, we study the widely used Mallows mixture model. In the high-dimensional setting, we propose a polynomial-time algorithm that learns a Mallows mixture of permutations on \( n \) elements with the optimal sample complexity that is proportional to \( \log n \), improving upon previous results that scale polynomially with \( n \). In the high-noise regime, we characterize the optimal dependency of the sample complexity on the noise parameter. Both objectives are accomplished by first studying demixing permutations under a noiseless query model using groups of pairwise comparisons, which can be viewed as moments of the mixing distribution, and then extending these results to the noisy Mallows model by simulating the noiseless oracle.

REFERENCES


Key words and phrases. Mixture of permutations, Mallows model, rank aggregation, group of pairwise comparisons, method of moments.


GENERALIZED RESILIENCE AND ROBUST STATISTICS

BY BANGHUA ZHU\textsuperscript{1,a}, JIANTAO JIAO\textsuperscript{1,b} AND JACOB STEINHARDT\textsuperscript{2,c}

\textsuperscript{1}Department of EECS, University of California, Berkeley, \textsuperscript{a}banghua@berkeley.edu, \textsuperscript{b}jiantao@berkeley.edu
\textsuperscript{2}Department of Statistics, University of California, Berkeley, \textsuperscript{c}jsteinhardt@berkeley.edu

Robust statistics traditionally focuses on outliers, or perturbations in total variation distance. However, a dataset could be maliciously corrupted in many other ways, such as systematic measurement errors and missing covariates. We consider corruption in either TV or Wasserstein distance, and show that robust estimation is possible whenever the true population distribution satisfies a property called \textit{generalized resilience}, which holds under moment or hypercontractive conditions. For TV corruption model, our finite-sample analysis improves over previous results for mean estimation with bounded $k$th moment, linear regression, and joint mean and covariance estimation. For $W_1$ corruption, we provide the first finite-sample guarantees for second moment estimation and linear regression.

Technically, our robust estimators are a generalization of minimum distance (MD) functionals, which project the corrupted distribution onto a given set of well-behaved distributions. The error of these MD functionals is bounded by a certain modulus of continuity, and we provide a systematic method for upper bounding this modulus for the class of generalized resilient distributions, which usually gives sharp population-level results and good finite-sample guarantees.

REFERENCES


MSC2020 subject classifications. Primary 62F35; secondary 62G35.

Key words and phrases. Robust statistics, minimum distance functional, total variation distance perturbation, Wasserstein distance perturbation.
NONREGULAR AND MINIMAX ESTIMATION OF INDIVIDUALIZED THRESHOLDS IN HIGH DIMENSION WITH BINARY RESPONSES

BY HUIJIE FENG\textsuperscript{1,a}, YANG NING\textsuperscript{1,b} AND JIWEI ZHAO\textsuperscript{2,c}

\textsuperscript{1}Department of Statistics and Data Science, Cornell University, \textsuperscript{a}hf279@cornell.edu, \textsuperscript{b}yn265@cornell.edu
\textsuperscript{2}Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, \textsuperscript{c}jiwei.zhao@wisc.edu

Given a large number of covariates $Z$, we consider the estimation of a high-dimensional parameter $\theta$ in an individualized linear threshold $\theta^T Z$ for a continuous variable $X$, which minimizes the disagreement between \text{sign}(X - \theta^T Z) and a binary response $Y$. While the problem can be formulated into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to discontinuity of the sign function. Moreover, estimating $\theta$ even in the fixed-dimensional setting is known as a nonregular problem leading to nonstandard asymptotic theory. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter $\theta$, we propose an empirical risk minimization approach based on a regularized smoothed non-convex loss function. The Fisher consistency of the proposed method is guaranteed as the bandwidth of the smoothed loss is shrunk to 0. Statistically, we show that the finite sample error bound for estimating $\theta$ in $\ell_2$ norm is $(s \log d/n)^{\beta/(2\beta+1)}$, where $d$ is the dimension of $\theta$, $s$ is the sparsity level, $n$ is the sample size and $\beta$ is the smoothness of the conditional density of $X$ given the response $Y$ and the covariates $Z$. The convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. The Lepski’s method is developed to achieve the adaption to the unknown sparsity $s$ or smoothness $\beta$. Computationally, an efficient path-following algorithm is proposed to compute the solution path. We show that this algorithm achieves geometric rate of convergence for computing the whole path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis from the ChAMP (Chondral Lesions And Meniscus Procedures) Trial.

REFERENCES


Key words and phrases. High-dimensional statistics, nonstandard asymptotics, non-convex optimization, minimax optimality, adaptivity, kernel method.


ON THE ROBUSTNESS OF MINIMUM NORM INTERPOLATORS AND REGULARIZED EMPIRICAL RISK MINIMIZERS

BY GEOFFREY CHINOT\(^a\), MATTHIAS LÖFFLER\(^b\) AND SARA VAN DE GEER\(^c\)

Seminar for Statistics, Department of Mathematics, ETH Zürich, \(^a\)geoffrey.chinot@stat.math.ethz.ch, \(^b\)matthias.loeffler@stat.math.ethz.ch, \(^c\)sara.vandegeer@stat.math.ethz.ch

This article develops a general theory for minimum norm interpolating estimators and regularized empirical risk minimizers (RERM) in linear models in the presence of additive, potentially adversarial, errors. In particular, no conditions on the errors are imposed. A quantitative bound for the prediction error is given, relating it to the Rademacher complexity of the covariates, the norm of the minimum norm interpolator of the errors and the size of the subdifferential around the true parameter.

The general theory is illustrated for Gaussian features and several norms: The \(\ell_1\), \(\ell_2\), group Lasso and nuclear norms. In case of sparsity or low-rank inducing norms, minimum norm interpolators and RERM yield a prediction error of the order of the average noise level, provided that the overparameterization is at least a logarithmic factor larger than the number of samples and that, in case of RERM, the regularization parameter is small enough.

Lower bounds that show near optimality of the results complement the analysis.

REFERENCES


MSC2020 subject classifications. Primary 62J05; secondary 65F45.

Key words and phrases. Sparse linear regression, regularization, basis pursuit, trace regression, interpolation, minimum norm interpolation.


RANDOM GRAPH ASYMPTOTICS FOR TREATMENT EFFECT ESTIMATION UNDER NETWORK INTERFERENCE

BY SHUANGNING LI\textsuperscript{1,a} AND STEFAN WAGER\textsuperscript{2,b}

\textsuperscript{1}Department of Statistics, Stanford University, \textsuperscript{a}lsn@stanford.edu
\textsuperscript{2}Graduate School of Business, Stanford University, \textsuperscript{b}swager@stanford.edu

The network interference model for treatment effect estimation places experimental units at the vertices of an undirected exposure graph, such that treatment assigned to one unit may affect the outcome of another unit if and only if these two units are connected by an edge. This model has recently gained popularity as means of incorporating interference effects into the Neyman–Rubin potential outcomes framework; and several authors have considered estimation of various causal targets, including the direct and indirect effects of treatment. In this paper, we consider large-sample asymptotics for treatment effect estimation under network interference in a setting where the exposure graph is a random draw from a graphon. When targeting the direct effect, we establish a central limit theorem and find that—in our setting—popular estimators are considerably more accurate than existing results suggest. Meanwhile, when targeting the indirect effect, we leverage our generative assumptions to propose a consistent estimator in a setting where no other consistent estimators are currently available. Overall, our results highlight the promise of random graph asymptotics in understanding the practicality and limits of causal inference under network interference.

REFERENCES


\textsuperscript{MSC2020 subject classifications.} 62G20.

\textsuperscript{Key words and phrases.} Causal inference, direct and indirect effects, graphon, potential outcome.
AN $\ell_p$ THEORY OF PCA AND SPECTRAL CLUSTERING

BY EMMANUEL ABBE$^{1,a}$, JIANQING FAN$^{2,b}$ AND KAIZHENG WANG$^{3,c}$

$^{1}$Institute of Mathematics, EPFL, $^a$emmanuel.abbe@epfl.ch
$^{2}$Department of ORFE, Princeton University, $^b$qfan@princeton.edu
$^{3}$Department of IEOR and Data Science Institute, Columbia University, $^c$kaizheng.wang@columbia.edu

Principal Component Analysis (PCA) is a powerful tool in statistics and machine learning. While existing study of PCA focuses on the recovery of principal components and their associated eigenvalues, there are few precise characterizations of individual principal component scores that yield low-dimensional embedding of samples. That hinders the analysis of various spectral methods. In this paper, we first develop an $\ell_p$ perturbation theory for a hollowed version of PCA in Hilbert spaces which provably improves upon the vanilla PCA in the presence of heteroscedastic noises. Through a novel $\ell_p$ analysis of eigenvectors, we investigate entrywise behaviors of principal component score vectors and show that they can be approximated by linear functionals of the Gram matrix in $\ell_p$ norm, which includes $\ell_2$ and $\ell_\infty$ as special cases. For sub-Gaussian mixture models, the choice of $p$ giving optimal bounds depends on the signal-to-noise ratio, which further yields optimality guarantees for spectral clustering. For contextual community detection, the $\ell_p$ theory leads to simple spectral algorithms that achieve the information threshold for exact recovery and the optimal misclassification rate.

REFERENCES


ESTIMATION OF SMOOTH FUNCTIONALS IN HIGH-DIMENSIONAL MODELS: BOOTSTRAP CHAINS AND GAUSSIAN APPROXIMATION

BY VLADIMIR KOLTCHINSKI\textsuperscript{a}

School of Mathematics, Georgia Institute of Technology, \textsuperscript{a}vlad@math.gatech.edu

Let $X^{(n)}$ be an observation sampled from a distribution $P^{(n)}_{\theta}$ with an unknown parameter $\theta$, $\theta$ being a vector in a Banach space $E$ (most often, a high-dimensional space of dimension $d$). We study the problem of estimation of $f(\theta)$ for a functional $f: E \mapsto \mathbb{R}$ of some smoothness $s > 0$ based on an observation $X^{(n)} \sim P^{(n)}_{\theta}$. Assuming that there exists an estimator $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ of parameter $\theta$ such that $\sqrt{n}(\hat{\theta}_n - \theta)$ is sufficiently close in distribution to a mean zero Gaussian random vector in $E$, we construct a functional $g: E \mapsto \mathbb{R}$ such that $g(\hat{\theta}_n)$ is an asymptotically normal estimator of $f(\theta)$ with $\sqrt{n}$ rate provided that $s > \frac{1}{1 - \alpha}$ and $d \leq n^\alpha$ for some $\alpha \in (0, 1)$. We also derive general upper bounds on Orlicz norm error rates for estimator $g(\hat{\theta})$ depending on smoothness $s$, dimension $d$, sample size $n$ and the accuracy of normal approximation of $\sqrt{n}(\hat{\theta}_n - \theta)$. In particular, this approach yields asymptotically efficient estimators in high-dimensional log-concave exponential models.

REFERENCES


Key words and phrases. Efficiency, smooth functionals, bootstrap chain, concentration inequalities, normal approximation.


A MINIMAX FRAMEWORK FOR QUANTIFYING RISK-FAIRNESS
TRADE-OFF IN REGRESSION

BY EVGENII CHZHEN¹,a AND NICOLAS SCHREUER¹,b

¹Laboratoire de mathématiques d’Orsay, CNRS, Université Paris-Saclay, aevgenii.chzhen@cnrs.fr
²DIBRIS, MaLGa, Università di Genova, bschreuder.nicolas@gmail.com

We propose a theoretical framework for the problem of learning a real-valued function which meets fairness requirements. This framework is built upon the notion of $\alpha$-relative (fairness) improvement of the regression function which we introduce using the theory of optimal transport. Setting $\alpha = 0$ corresponds to the regression problem under the Demographic Parity constraint, while $\alpha = 1$ corresponds to the classical regression problem without any constraints. For $\alpha \in (0, 1)$ the proposed framework allows to continuously interpolate between these two extreme cases and to study partially fair predictors. Within this framework, we precisely quantify the cost in risk induced by the introduction of the fairness constraint. We put forward a statistical minimax setup and derive a general problem-dependent lower bound on the risk of any estimator satisfying $\alpha$-relative improvement constraint. We illustrate our framework on a model of linear regression with Gaussian design and systematic group-dependent bias, deriving matching (up to absolute constants) upper and lower bounds on the minimax risk under the introduced constraint. We provide a general post-processing strategy which enjoys fairness, risk guarantees and can be applied on top of any black-box algorithm. Finally, we perform a simulation study of the linear model and numerical experiments of benchmark data, validating our theoretical contributions.

REFERENCES


Key words and phrases. Algorithmic fairness, risk-fairness trade-off, regressions, demographic parity, least-squares, optimal transport, minimax analysis, statistical learning, Wasserstein barycenter, optimal transport, Pareto optimality.


CONSISTENCY OF INVARIANCE-BASED RANDOMIZATION TESTS

BY EDGAR DOBRIKAN

Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, 

dobriban@wharton.upenn.edu

Invariance-based randomization tests—such as permutation tests, rotation tests, or sign changes—are an important and widely used class of statistical methods. They allow drawing inferences under weak assumptions on the data distribution. Most work focuses on their type I error control properties, while their consistency properties are much less understood.

We develop a general framework to study the consistency of invariance-based randomization tests, assuming the data is drawn from a signal-plus-noise model. We allow the transforms (e.g., permutations or rotations) to be general compact topological groups, such as rotation groups, acting by linear group representations. We study test statistics with a generalized subadditivity property.

We apply our framework to a number of fundamental and highly important problems in statistics, including sparse vector detection, testing for low-rank matrices in noise, sparse detection in linear regression, and two-sample testing. Comparing with minimax lower bounds we develop, we find perhaps surprisingly that in some cases, randomization tests detect signals at the minimax optimal rate.

REFERENCES


Key words and phrases. Randomization test, permutation test, nonparametric group invariance, sparse detection.


ERRATUM: ASYMPTOTIC GENEALOGIES OF INTERACTING PARTICLE SYSTEMS WITH AN APPLICATION TO SEQUENTIAL MONTE CARLO

BY JERE KOSKELA\textsuperscript{a}, PAUL A. JENKINS\textsuperscript{b}, ADAM M. JOHANSEN\textsuperscript{c} AND DARIO SPANÒ\textsuperscript{d}

Department of Statistics, University of Warwick, \textsuperscript{a}j.koskela@warwick.ac.uk, \textsuperscript{b}p.jenkins@warwick.ac.uk, \textsuperscript{c}a.m.johansen@warwick.ac.uk, \textsuperscript{d}d.spano@warwick.ac.uk

REFERENCES


Key words and phrases. Coalescent, finite-dimensional distributions, genealogy, interacting particle system, sequential Monte Carlo.