The Annals of Statistics

An Official Journal of the Institute of Mathematical Statistics

Articles

Orthogonal statistical learning.............Dylan J. Foster and Vasilis Syrgkanis 879
Optimal high-dimensional and nonparametric distributed testing under communication constraints........Botond Szabó, Lasse Vuursteen and Harry van Zanten 909
Optimal Permutation Estimation in CrowdSourcing problems Emmanuel Pilliat, Alexandra Carpentier and Nicolas Verzele 935
Total positivity in multivariate extremes Frank Röttger, Sebastian Engelke and Piotr Zwiernik 962
A power analysis for model-X knockoffs with $\ell_p$-regularized statistics Asaf Weinstein, Weiji Su, Małgorzata Bogdan, Rina Foygel Barber and Emmanuel J. Candes 1005
A general characterization of optimal tie-breaker designs Harrison H. Li and Art B. Owen 1030
Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes........Quan Zhou and Hyunwoong Chang 1058
Optimal reach estimation and metric learning Eddy Aamari, Clément Berenfeld and Clément Levrard 1086
Universal rank inference via residual subsampling with application to large networks Xiao Han, Qing Yang and Yingying Fan 1109
Pairwise interaction function estimation of stationary Gibbs point processes using basis expansion.......................Ismaila Ba, Jean-François Coeurjolly and Francisco Cuevas-Pacheco 1134
Efficiency of estimators for locally asymptotically normal quantum statistical models Akio Fujihara and Koichi Yamagata 1159
Inference in Ising models on dense regular graphs Yuanzhe Xu and Sumit Mukherjee 1183
AutoRegressive approximations to nonstationary time series with inference and applications.......................Xiucai Ding and Zhou Zhou 1207
Optimal discriminant analysis in high-dimensional latent factor models Xin Bing and Marten Wegkamp 1232
Dispersal density estimation across scales....Marc Hoffmann and Mathias Trabs 1258
Minimax rate of distribution estimation on unknown submanifolds under adversarial losses.........................Rong Tang and Yun Yang 1282
Inference for low-rank models Victor Chernozhukov, Christian Hansen, Yuan Liao and Yinchu Zhu 1309
Extreme value inference for heterogeneous power law data John H. J. Einmahl and Yi He 1331
Finite-sample complexity of sequential Monte Carlo estimators Joe Marion, Joseph Mathews and Scott C. Schmidler 1357
Coverage of credible intervals in Bayesian multivariate isotonic regression Kang Wang and Subhashis Ghosal 1376
Universal regression with adversarial responses Moïse Blanchard and Patrick Jaillet 1401
Asymptotic normality for eigenvalue statistics of a general sample covariance matrix when $p/n \to \infty$ and applications........Jiaxin Qiu, Zeng Li and Jianfeng Yao 1427
ORTHOGONAL STATISTICAL LEARNING

BY DYLAN J. FOSTER\textsuperscript{1,a} AND VASILIS SYRGKANIS\textsuperscript{2,b}

\textsuperscript{1}Microsoft Research, New England, \textsuperscript{a}dylanfoster@microsoft.com
\textsuperscript{2}Department of Management Science and Engineering, Stanford University, \textsuperscript{b}vsyrgk@stanford.edu

We provide nonasymptotic excess risk guarantees for statistical learning in a setting where the population risk with respect to which we evaluate the target parameter depends on an unknown \textit{nuisance parameter} that must be estimated from data. We analyze a two-stage sample splitting meta-algorithm that takes as input arbitrary estimation algorithms for the target parameter and nuisance parameter. We show that if the population risk satisfies a condition called \textit{Neyman orthogonality}, the impact of the nuisance estimation error on the excess risk bound achieved by the meta-algorithm is of second order. Our theorem is agnostic to the particular algorithms used for the target and nuisance and only makes an assumption on their individual performance. This enables the use of a plethora of existing results from machine learning to give new guarantees for learning with a nuisance component. Moreover, by focusing on excess risk rather than parameter estimation, we can provide rates under weaker assumptions than in previous works and accommodate settings in which the target parameter belongs to a complex nonparametric class. We provide conditions on the metric entropy of the nuisance and target classes such that \textit{oracle rates} of the same order, as if we knew the nuisance parameter, are achieved.

REFERENCES


\textit{MSC2020 subject classifications}. Primary 62G08; secondary 62C20, 62D20.

\textit{Key words and phrases}. Statistical learning, double machine learning, Neyman orthogonality, policy learning, treatment effects, local Rademacher complexity.


OPTIMAL HIGH-DIMENSIONAL AND NONPARAMETRIC DISTRIBUTED TESTING UNDER COMMUNICATION CONSTRAINTS

BY BOTOND SZABÓ¹,a, LASSE VUURSTEEN²,b AND HARRY VAN ZANTEN³,c

¹Department of Decision Sciences, Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University,  
²Delft Institute of Applied Mathematics (DIAM), Delft University of Technology,  
³Department of Mathematics, Vrije Universiteit Amsterdam, 
botszabo@unibocconi.it

We derive minimax testing errors in a distributed framework where the data is split over multiple machines and their communication to a central machine is limited to $b$ bits. We investigate both the $d$- and infinite-dimensional signal detection problem under Gaussian white noise. We also derive distributed testing algorithms reaching the theoretical lower bounds.

Our results show that distributed testing is subject to fundamentally different phenomena that are not observed in distributed estimation. Among our findings we show that testing protocols that have access to shared randomness can perform strictly better in some regimes than those that do not. We also observe that consistent nonparametric distributed testing is always possible, even with as little as one bit of communication, and the corresponding test outperforms the best local test using only the information available at a single local machine. Furthermore, we also derive adaptive nonparametric distributed testing strategies and the corresponding theoretical lower bounds.

REFERENCES


MSC2020 subject classifications. Primary 62G10, 62F30; secondary 62F03.

Key words and phrases. Distributed methods, nonparametric, hypothesis testing, minimax optimal.


OPTIMAL PERMUTATION ESTIMATION IN CROWDSOURCING PROBLEMS

BY EMMANUEL PILLIAT\textsuperscript{1,}a, ALEXANDRA CARPENTIER\textsuperscript{2,}b AND NICOLAS VERZELEN\textsuperscript{3,}c

\textsuperscript{1}IMAG, Université Montpellier, \textsuperscript{a}emmanuel.pilliat@umontpellier.fr
\textsuperscript{2}Institut für Mathematik, Universität Potsdam, \textsuperscript{b}carpentier@uni-potsdam.de
\textsuperscript{3}MISTEA, INRAE, Université Montpellier, \textsuperscript{c}nicolas.verzelen@inrae.fr

Motivated by crowdsourcing applications, we consider a model where we have partial observations from a bivariate isotonic $n \times d$ matrix with an unknown permutation $\pi^*$ acting on its rows. Focusing on the twin problems of recovering the permutation $\pi^*$ and estimating the unknown matrix, we introduce a polynomial-time procedure achieving the minimax risk for these two problems, this for all possible values of $n$, $d$, and all possible sampling efforts. Along the way we establish that, in some regimes, recovering the unknown permutation $\pi^*$ is considerably simpler than estimating the matrix.

REFERENCES


MSC2020 subject classifications. 62C20.

Key words and phrases. Bi-isotonic matrices, ranking, sub-Gaussian noise.


TOTAL POSITIVITY IN MULTIVARIATE EXTREMES

BY FRANK RÖTTGER\textsuperscript{1,a}, SEBASTIAN ENGELKE\textsuperscript{1,b} AND PIOTR ZWIERNIK\textsuperscript{2,c}

\textsuperscript{1}Research Center for Statistics, Université de Genève, \textsuperscript{a}frank.roettger@unige.ch, \textsuperscript{b}sebastian.engelke@unige.ch
\textsuperscript{2}Department of Statistical Sciences, University of Toronto, \textsuperscript{c}piotr.zwiernik@utoronto.ca

Positive dependence is present in many real world data sets and has appealing stochastic properties that can be exploited in statistical modeling and in estimation. In particular, the notion of multivariate total positivity of order 2 (MTP\textsubscript{2}) is a convex constraint and acts as an implicit regularizer in the Gaussian case. We study positive dependence in multivariate extremes and introduce EMTP\textsubscript{2}, an extremal version of MTP\textsubscript{2}. This notion turns out to appear prominently in extremes, and in fact, it is satisfied by many classical models. For a Hüsler–Reiss distribution, the analogue of a Gaussian distribution in extremes, we show that it is EMTP\textsubscript{2} if and only if its precision matrix is a Laplacian of a connected graph. We propose an estimator for the parameters of the Hüsler–Reiss distribution under EMTP\textsubscript{2} as the solution of a convex optimization problem with Laplacian constraint. We prove that this estimator is consistent and typically yields a sparse model with possibly non-decomposable extremal graphical structure. Applying our methods to a data set of Danube River flows, we illustrate this regularization and the superior performance compared to existing methods.

REFERENCES


MSC2020 subject classifications. Primary 60G70, 62H22; secondary 15B48.
Key words and phrases. Convex optimization, extreme value theory, covariance mapping, graph Laplacians, total positivity.
A POWER ANALYSIS FOR MODEL-X KNOCKOFFS WITH ℓ_p-REGULARIZED STATISTICS

BY ASAF WEINSTEIN1,a, WEIJIE J. SU2,b, MAŁGORZATA BOGDAN3,c, RINA FOYGEL BARBER4,d AND EMMANUEL J. CANDÈS5,e

1Department of Statistics and Data Science, Hebrew University of Jerusalem, asaf.weinstein@mail.huji.ac.il
2Department of Statistics and Data Science, University of Pennsylvania, suw@wharton.upenn.edu
3Department of Mathematics, University of Wroclaw, malgorzata.bogdan@uw.edu.pl
4Department of Statistics, University of Chicago, rina@uchicago.edu
5Department of Statistics and Department of Mathematics, Stanford University, candes@stanford.edu

Variable selection properties of procedures utilizing penalized-likelihood estimates is a central topic in the study of high-dimensional linear regression problems. Existing literature emphasizes the quality of ranking of the variables by such procedures as reflected in the receiver operating characteristic curve or in prediction performance. Specifically, recent works have harnessed modern theory of approximate message-passing (AMP) to obtain, in a particular setting, exact asymptotic predictions of the type I, type II error tradeoff for selection procedures that rely on ℓ_p-regularized estimators.

In practice, effective ranking by itself is often not sufficient because some calibration for Type I error is required. In this work, we study theoretically the power of selection procedures that similarly rank the features by the size of an ℓ_p-regularized estimator, but further use Model-X knockoffs to control the false discovery rate in the realistic situation where no prior information about the signal is available. In analyzing the power of the resulting procedure, we extend existing results in AMP theory to handle the pairing between original variables and their knockoffs. This is used to derive exact asymptotic predictions for power. We apply the general results to compare the power of the knockoffs versions of Lasso and thresholded-Lasso selection, and demonstrate that in the i.i.d. covariate setting under consideration, tuning by cross-validation on the augmented design matrix is nearly optimal. We further demonstrate how the techniques allow to analyze also the Type S error, and a corresponding notion of power, when selections are supplemented with a decision on the sign of the coefficient.

REFERENCES


MSC2020 subject classifications. 62F03, 62F07, 62F05, 62F12.
Key words and phrases. Knockoffs, variable selection, false discovery rate (FDR), power analysis, high-dimensional regression.


A GENERAL CHARACTERIZATION OF OPTIMAL TIE-BREAKER DESIGNS

BY HARRISON H. LI\textsuperscript{a} AND ART B. OWEN\textsuperscript{b}

Department of Statistics, Stanford University, \textsuperscript{a}hli90722@stanford.edu, \textsuperscript{b}owen@stanford.edu

Tie-breaker designs trade off a measure of statistical efficiency against a short-term gain from preferentially assigning a binary treatment to subjects with higher values of a running variable $x$. The efficiency measure can be any continuous function of the expected information matrix in a two-line regression model. The short-term gain is expressed as the covariance between the running variable and the treatment indicator. We investigate how to choose design functions $p(x)$ specifying the probability of treating a subject with running variable $x$ in order to optimize these competing objectives, under external constraints on the number of subjects receiving treatment. Our results include sharp existence and uniqueness guarantees, while accommodating the ethically appealing requirement that $p(x)$ be nondecreasing in $x$. Under this condition, there is always an optimal treatment probability function $p(x)$ that is constant on the sets $(-\infty, t)$ and $(t, \infty)$ for some threshold $t$ and generally discontinuous at $x = t$. When the running variable distribution is not symmetric or the fraction of subjects receiving the treatment is not $1/2$, our optimal designs improve upon a $D$-optimality objective without sacrificing short-term gain, compared to a typical three-level tie-breaker design that fixes treatment probabilities at 0, 1/2 and 1. We illustrate our optimal designs with data from Head Start, an early childhood government intervention program.

REFERENCES


COMPLEXITY ANALYSIS OF BAYESIAN LEARNING OF HIGH-DIMENSIONAL DAG MODELS AND THEIR EQUIVALENCE CLASSES

BY QUAN ZHOU and HYUNWOOONG CHANG

Department of Statistics, Texas A&M University, a quan@stat.tamu.edu, b hwchang@stat.tamu.edu

Structure learning via MCMC sampling is known to be very challenging because of the enormous search space and the existence of Markov equivalent DAGs. Theoretical results on the mixing behavior are lacking. In this work, we prove the rapid mixing of a random walk Metropolis–Hastings algorithm, which reveals that the complexity of Bayesian learning of sparse equivalence classes grows only polynomially in \( n \) and \( p \), under some high-dimensional assumptions. A series of high-dimensional consistency results is obtained, including the strong selection consistency of an empirical Bayes model for structure learning. Our proof is based on two new results. First, we derive a general mixing time bound on finite-state spaces, which can be applied to local MCMC schemes for other model selection problems. Second, we construct high-probability search paths on the space of equivalence classes with node degree constraints by proving a combinatorial property of DAG comparisons. Simulation studies on the proposed MCMC sampler are conducted to illustrate the main theoretical findings.

REFERENCES


MSC2020 subject classifications. 62F15, 62J05.
Key words and phrases. Finite Markov chains, greedy equivalence search (GES), locally informed proposals, Poincaré-type inequality, random walk Metropolis–Hastings, rapid mixing, strong selection consistency.


OPTIMAL REACH ESTIMATION AND METRIC LEARNING

BY EDDIE AAMARI¹,a, CLÉMENT BERENFELD²,c AND CLÉMENT LEVRARD¹,b

¹LPSM–CNRS, Université Paris Cité, Sorbonne Université, a.aamari@lpsm.paris, b.levrard@lpsm.paris
²Institut für Mathematik, Universität Potsdam, c.berenfeld@uni-potsdam.de

We study the estimation of the reach, an ubiquitous regularity parameter in manifold estimation and geometric data analysis. Given an i.i.d. sample over an unknown $d$-dimensional $C^k$-smooth submanifold $M$ of $\mathbb{R}^D$, we provide optimal nonasymptotic bounds for the estimation of its reach. We build upon a formulation of the reach in terms of maximal curvature on one hand and geodesic metric distortion on the other. The derived rates are adaptive, with rates depending on whether the reach of $M$ arises from curvature or from a bottleneck structure. In the process we derive optimal geodesic metric estimation bounds.

REFERENCES


MSC2020 subject classifications. 62G05, 62C20, 68U05.
Key words and phrases. Geometric inference, reach, metric learning, manifold learning, minimax risk.


UNIVERSAL RANK INFERENCE VIA RESIDUAL SUBSAMPLING WITH APPLICATION TO LARGE NETWORKS

BY XIAO HAN\textsuperscript{1,a}, QING YANG\textsuperscript{1,b} AND YINGYING FAN\textsuperscript{2,c}

\textsuperscript{1}International Institute of Finance, School of Management, University of Science and Technology of China, \textsuperscript{a}xhan011@ustc.edu.cn, \textsuperscript{b}yangq@ustc.edu.cn

\textsuperscript{2}Data Sciences and Operations Department, University of Southern California, \textsuperscript{c}fanyingy@marshall.usc.edu

Determining the precise rank is an important problem in many large-scale applications with matrix data exploiting low-rank plus noise models. In this paper, we suggest a universal approach to rank inference via residual subsampling (RIRS) for testing and estimating rank in a wide family of models, including many popularly used network models such as the degree corrected mixed membership model as a special case. Our procedure constructs a test statistic via subsampling entries of the residual matrix after extracting the spiked components. The test statistic converges in distribution to the standard normal under the null hypothesis, and diverges to infinity with asymptotic probability one under the alternative hypothesis. The effectiveness of RIRS procedure is justified theoretically, utilizing the asymptotic expansions of eigenvectors and eigenvalues for large random matrices recently developed in (J. Amer. Statist. Assoc. 117 (2022) 996–1009) and (J. R. Stat. Soc. Ser. B. Stat. Methodol. 84 (2022) 630–653). The advantages of the newly suggested procedure are demonstrated through several simulation and real data examples.

REFERENCES


PAIRWISE INTERACTION FUNCTION ESTIMATION OF STATIONARY GIBBS POINT PROCESSES USING BASIS EXPANSION

BY ISMAÏLA BA¹,a, JEAN-FRANÇOIS COEURJOLLY²,b AND FRANCISCO CUEVAS-PACHECO³,c

¹Department of Mathematics, Université du Québec à Montréal (UQAM), a ba.ismaila@courrier.uqam.ca
²Laboratory Jean Kuntzmann, Department DATA, Univ. Grenoble Alpes, b jean-francois.coeurjolly@univ-grenoble-alpes.fr
³Departamento de Matemática, Universidad Técnica Federico Santa María, c francisco.cuevas@usm.cl

The class of Gibbs point processes (GPP) is a large class of spatial point processes able to model both clustered and repulsive point patterns. They are specified by their conditional intensity, which for a point pattern \( x \) and a location \( u \), is roughly speaking the probability that an event occurs in an infinitesimal ball around \( u \) given the rest of the configuration is \( x \). The most simple and natural class of models is the class of pairwise interaction point processes where the conditional intensity depends on the number of points and pairwise distances between them. This paper is concerned with the problem of estimating the pairwise interaction function nonparametrically. We propose to estimate it using an orthogonal series expansion of its logarithm. Such an approach has numerous advantages compared to existing ones. The estimation procedure is simple, fast and completely data-driven. We provide asymptotic properties such as consistency and asymptotic normality and show the efficiency of the procedure through simulation experiments and illustrate it with several data sets.

REFERENCES


MSC2020 subject classifications. Primary62H11, 60G55; secondary62J07, 65C60, 97K80.

Key words and phrases. Spatial statistics, Gibbs point process, orthogonal series estimator, pairwise interaction function.


EFFICIENCY OF ESTIMATORS FOR LOCALLY ASYMPTOTICALLY NORMAL QUANTUM STATISTICAL MODELS

BY AKIO FUJIWARA1,a AND KOICHI YAMAGATA2,b

1Department of Mathematics, Osaka University, afujiwara@math.sci.osaka-u.ac.jp
2Principles of Informatics Research Division, National Institute of Informatics, byamagata@nii.ac.jp

We herein establish an asymptotic representation theorem for locally asymptotically normal quantum statistical models. This theorem enables us to study the asymptotic efficiency of quantum estimators, such as quantum regular estimators and quantum minimax estimators, leading to a universal tight lower bound beyond the i.i.d. assumption. This formulation complements the theory of quantum contiguity developed in the previous paper [Fujiwara and Yamagata, Bernoulli 26 (2020) 2105–2141], providing a solid foundation of the theory of weak quantum local asymptotic normality.

REFERENCES


MSC2020 subject classifications. Primary 62F12, 81P50; secondary 62A01.

Key words and phrases. Quantum statistics, local asymptotic normality, asymptotic representation theorem, asymptotic efficiency, regular estimator, minimax estimator.


INFECTION IN ISING MODELS ON DENSE REGULAR GRAPHS

BY YUANZHE XUA AND SUMIT MUKHERJEEB

Department of Statistics, Columbia University, a yuanzhe.xu@columbia.edu, b sm3949@columbia.edu

In this paper, we derive the limit of experiments for one-parameter Ising models on dense regular graphs. In particular, we show that the limiting experiment is Gaussian in the “low temperature” regime, and non-Gaussian in the “critical” regime. We also derive the limiting distributions of the maximum likelihood and maximum pseudolikelihood estimators, and study limiting power for tests of hypothesis against contiguous alternatives. To the best of our knowledge, this is the first attempt at establishing the classical limits of experiments for Ising models (and more generally, Markov random fields).

REFERENCES


Key words and phrases. Ising model, limits of experiments, phase transition, asymptotic power, asymptotic efficiency.
Understanding the time-varying structure of complex temporal systems is one of the main challenges of modern time-series analysis. In this paper, we show that every uniformly-positive-definite-in-covariance and sufficiently short-range dependent nonstationary and nonlinear time series can be well approximated globally by a white-noise-driven autoregressive (AR) process of slowly diverging order. To our best knowledge, it is the first time such a structural approximation result is established for general classes of nonstationary time series. A high-dimensional \( L^2 \) test and an associated multiplier bootstrap procedure are proposed for the inference of the AR approximation coefficients. In particular, an adaptive stability test is proposed to check whether the AR approximation coefficients are time-varying, a frequently encountered question for practitioners and researchers of time series. As an application, globally optimal sffollowing hort-term forecasting theory and methodology for a wide class of locally stationary time series are established via the method of sieves.

REFERENCES


MSC2020 subject classifications. Primary 62M10, 62M20; secondary 60G07.
Key words and phrases. Nonstationary time series, AR approximation, high-dimensional convex Gaussian approximation, multiplier bootstrap, globally optimal forecasting.
PAPARODITIS, E. (2010). Validating stationarity assumptions in time series analysis by rolling local pe-
MAYER, U., ZÄHLE, H. and ZHOU, Z. (2020). Functional weak limit theorem for a local empiri-
POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Uncon-


OPTIMAL DISCRIMINANT ANALYSIS IN HIGH-DIMENSIONAL LATENT FACTOR MODELS

BY XIN BING\(^1\),\(^a\) AND MARTEN WEGKAMP\(^2\),\(^b\)

\(^1\)Department of Statistical Sciences, University of Toronto, \(^a\)xin.bing@utoronto.ca

\(^2\)Department of Mathematics & Department of Statistics and Data Science, Cornell University, \(^b\)mhw73@cornell.edu

In high-dimensional classification problems, a commonly used approach is to first project the high-dimensional features into a lower-dimensional space, and base the classification on the resulting lower-dimensional projections. In this paper, we formulate a latent-variable model with a hidden low-dimensional structure to justify this two-step procedure and to guide which projection to choose. We propose a computationally efficient classifier that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. A general theory is established for analyzing such two-step classifiers based on any projections. We derive explicit rates of convergence of the excess risk of the proposed PC-based classifier. The obtained rates are further shown to be optimal up to logarithmic factors in the minimax sense. Our theory allows the lower dimension to grow with the sample size and is also valid even when the feature dimension (greatly) exceeds the sample size. Extensive simulations corroborate our theoretical findings. The proposed method also performs favorably relative to other existing discriminant methods on three real data examples.

REFERENCES


MSC2020 subject classifications. 62H12, 62J07.

Key words and phrases. High-dimensional classification, latent factor model, principal component regression, dimension reduction, discriminant analysis, optimal rate of convergence.


We consider a space structured population model generated by two-point clouds: a homogeneous Poisson process $M$ with intensity $n \to \infty$ as a model for a parent generation together with a Cox point process $N$ as offspring generation, with conditional intensity given by the convolution of $M$ with a scaled dispersal density $\sigma^{-1} f (\cdot / \sigma)$. Based on a realisation of $M$ and $N$, we study the nonparametric estimation of $f$ and the estimation of the physical scale parameter $\sigma > 0$ simultaneously for all regimes $\sigma = \sigma_n$. We establish that the optimal rates of convergence do not depend monotonously on the scale and we construct minimax estimators accordingly whether $\sigma$ is known or considered as a nuisance, in which case we can estimate it and achieve asymptotic minimaxity by plug-in. The statistical reconstruction exhibits a competition between a direct and a deconvolution problem. Our study reveals in particular the existence of a least favorable intermediate inference scale, a phenomenon that seems to be new.

REFERENCES


MSC2020 subject classifications. Primary 62G05; secondary 62G07, 62M30, 60G57.

Key words and phrases. Nonparametric estimation and minimax theory, point processes, statistical inference across scales, dispersal models, deconvolution.
MINIMAX RATE OF DISTRIBUTION ESTIMATION ON UNKNOWN SUBMANIFOLDS UNDER ADVERSARIAL LOSSES

BY RONG TANG\textsuperscript{1,a} AND YUN YANG\textsuperscript{2,b}

\textsuperscript{1}Department of Mathematics, The Hong Kong University of Science and Technology, \textsuperscript{2}Department of Statistics, University of Illinois Urbana-Champaign

Statistical inference from high-dimensional data with low-dimensional structures has recently attracted a lot of attention. In machine learning, deep generative modelling approaches implicitly estimate distributions of complex objects by creating new samples from the underlying distribution, and have achieved great success in generating synthetic realistic-looking images and texts. A key step in these approaches is the extraction of latent features or representations (encoding) that can be used for accurately reconstructing the original data (decoding). In other words, low-dimensional manifold structure is implicitly assumed and utilized in the distribution modelling and estimation. To understand the benefit of low-dimensional manifold structure in generative modelling, we build a general minimax framework for distribution estimation on unknown submanifold under adversarial losses, with suitable smoothness assumptions on the target distribution and the manifold. The established minimax rate elucidates how various problem characteristics, including intrinsic dimensionality of the data and smoothness levels of the target distribution and the manifold, affect the fundamental limit of high-dimensional distribution estimation. To prove the minimax upper bound, we construct an estimator based on a mixture of locally fitted generative models, which is motivated by the partition of unity technique from differential geometry and is necessary to cover cases where the underlying data manifold does not admit a global parametrization. We also propose a data-driven adaptive estimator that is shown to simultaneously attain within a logarithmic factor of the optimal rate over a large collection of distribution classes.

REFERENCES


MSC2020 subject classifications. Primary 62C20, 62G07; secondary 62G05.

Key words and phrases. Adversarial training, generative model, distribution estimation, manifold, minimax rate, partition of unity.
This paper studies inference in linear models with a high-dimensional parameter matrix that can be well approximated by a “spiked low-rank matrix.” A spiked low-rank matrix has rank that grows slowly compared to its dimensions and nonzero singular values that diverge to infinity. We show that this framework covers a broad class of models of latent variables, which can accommodate matrix completion problems, factor models, varying coefficient models and heterogeneous treatment effects. For inference, we apply a procedure that relies on an initial nuclear-norm penalized estimation step followed by two ordinary least squares regressions. We consider the framework of estimating incoherent eigenvectors and use a rotation argument to argue that the eigenspace estimation is asymptotically unbiased. Using this framework, we show that our procedure provides asymptotically normal inference and achieves the semiparametric efficiency bound. We illustrate our framework by providing low-level conditions for its application in a treatment effects context where treatment assignment might be strongly dependent.

REFERENCES


Key words and phrases. Singular value thresholding, sample splitting, incoherent eigenvectors.


EXTREME VALUE INFERENCE FOR HETEROGENEOUS POWER LAW DATA

BY JOHN H.J. EINMAHL\textsuperscript{1,a} AND YI HE\textsuperscript{2,b}  

\textsuperscript{1}Department of Econometrics & OR and CentER, Tilburg University, \textsuperscript{a}john.h.j.einmahl@uvt.nl  
\textsuperscript{2}Amsterdam School of Economics, University of Amsterdam, \textsuperscript{b}yi.he@uva.nl

We extend extreme value statistics to independent data with possibly very different distributions. In particular, we present novel asymptotic normality results for the Hill estimator, which now estimates the extreme value index of the average distribution. Due to the heterogeneity, the asymptotic variance can be substantially smaller than that in the i.i.d. case. As a special case, we consider a heterogeneous scales model where the asymptotic variance can be calculated explicitly. The primary tool for the proofs is the functional central limit theorem for a weighted tail empirical process. We also present asymptotic normality results for the extreme quantile estimator. A simulation study shows the good finite-sample behavior of our limit theorems. We also present applications to assess the tail heaviness of earthquake energies and of cross-sectional stock market losses.

REFERENCES


\textbf{MSC2020 subject classifications.} Primary 62G32, 62G20, 62G30; secondary 60G70, 60F17.

\textbf{Key words and phrases.} Extreme value statistics, functional central limit theorem, heterogeneous scales model, Hill estimator, nonidentical distributions, weighted tail empirical process.


FINITE-SAMPLE COMPLEXITY OF SEQUENTIAL MONTE CARLO ESTIMATORS

BY JOE MARION\textsuperscript{a}, JOSEPH MATHEWS\textsuperscript{b} AND SCOTT C. SCHMIDLER\textsuperscript{c}

Department of Statistical Science, Duke University, \textsuperscript{a}joseph@berryconsultants.net, \textsuperscript{b}joseph.mathews@duke.edu, \textsuperscript{c}schmidler@stat.duke.edu

We present bounds for the finite-sample error of sequential Monte Carlo samplers on static spaces. Our approach explicitly relates the performance of the algorithm to properties of the chosen sequence of distributions and mixing properties of the associated Markov kernels. This allows us to give the first finite-sample comparison to other Monte Carlo schemes. We obtain bounds for the complexity of sequential Monte Carlo approximations for a variety of target distributions such as finite spaces, product measures and log-concave distributions including Bayesian logistic regression. The bounds obtained are within a logarithmic factor of similar bounds obtainable for Markov chain Monte Carlo.

REFERENCES


\textbf{MSC2020 subject classifications.} Primary 65C60; secondary 65C60, 60J22.

\textbf{Key words and phrases.} Sequential Monte Carlo, computational complexity, Bayesian computation.


We consider the nonparametric multivariate isotonic regression problem, where the regression function is assumed to be nondecreasing with respect to each predictor. Our goal is to construct a Bayesian credible interval for the function value at a given interior point with assured limiting frequentist coverage. A natural prior on the regression function is given by a random step function with a suitable prior on increasing step-heights, but the resulting posterior distribution is hard to analyze theoretically due to the complicated order restriction on the coefficients. We instead put a prior on unrestricted step-functions, but make inference using the induced posterior measure by an “immersion map” from the space of unrestricted functions to that of multivariate monotone functions. This allows for maintaining the natural conjugacy for posterior sampling. A natural immersion map to use is a projection with respect to a distance function, but in the present context, a block isotonization map is found to be more useful. The approach of using the induced “immersion posterior” measure instead of the original posterior to make inference provides a useful extension of the Bayesian paradigm, particularly helpful when the model space is restricted by some complex relations. We establish a key weak convergence result for the posterior distribution of the function at a point in terms of some functional of a multiindexed Gaussian process that leads to an expression for the limiting coverage of the Bayesian credible interval. Analogous to a recent result for univariate monotone functions, we find that the limiting coverage is slightly higher than the credibility, the opposite of a phenomenon observed in smoothing problems. Interestingly, the relation between credibility and limiting coverage does not involve any unknown parameter. Hence, by a recalibration procedure, we can get a predetermined asymptotic coverage by choosing a suitable credibility level smaller than the targeted coverage, and thus also shorten the credible intervals.

REFERENCES


Key words and phrases. Isotonic regression, credible interval, limiting coverage, Gaussian process, Block isotonization, immersion posterior.
UNIVERSAL REGRESSION WITH ADVERSARIAL RESPONSES

BY MOÏSE BLANCHARD\textsuperscript{1,a} AND PATRICK JAILLET\textsuperscript{2,b}

\textsuperscript{1}Operations Research Center, Massachusetts Institute of Technology, \textsuperscript{a}moiseb@mit.edu
\textsuperscript{2}Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, \textsuperscript{b}jaillet@mit.edu

We provide algorithms for regression with adversarial responses under large classes of non-i.i.d. instance sequences, on general separable metric spaces, with provably minimal assumptions. We also give characterizations of learnability in this regression context. We consider universal consistency, which asks for strong consistency of a learner without restrictions on the value responses. Our analysis shows that such an objective is achievable for a significantly larger class of instance sequences than stationary processes, and unveils a fundamental dichotomy between value spaces: whether finite-horizon mean estimation is achievable or not. We further provide optimistically universal learning rules, that is, such that if they fail to achieve universal consistency, any other algorithms will fail as well. For unbounded losses, we propose a mild integrability condition under which there exist algorithms for adversarial regression under large classes of non-i.i.d. instance sequences. In addition, our analysis also provides a learning rule for mean estimation in general metric spaces that is consistent under adversarial responses without any moment conditions on the sequence, a result of independent interest.

REFERENCES


MSC2020 subject classifications. Primary 62C05, 62J02, 97K80, 60G20; secondary 54E70.

Key words and phrases. Statistical learning theory, consistency, nonparametric estimation, generalization, stochastic processes, online learning, metric spaces.
ASYMPTOTIC NORMALITY FOR EIGENVALUE STATISTICS OF A GENERAL SAMPLE COVARIANCE MATRIX WHEN \( p/n \to \infty \) AND APPLICATIONS

BY JIAXIN QIU\(^1\), ZENG LI\(^2\) and JIANFENG YAO\(^3\)

\(^1\)Department of Statistics and Actuarial Science, The University of Hong Kong, \(^2\)Department of Statistics and Data Science, Southern University of Science and Technology, \(^3\)School of Data Science, The Chinese University of Hong Kong (Shenzhen)

The asymptotic normality for a large family of eigenvalue statistics of a general sample covariance matrix is derived under the ultrahigh-dimensional setting, that is, when the dimension to sample size ratio \( p/n \to \infty \). Based on this CLT result, we extend the covariance matrix test problem to the new ultra-high-dimensional context, and apply it to test a matrix-valued white noise. Simulation experiments are conducted for the investigation of finite-sample properties of the general asymptotic normality of eigenvalue statistics, as well as the two developed tests.

REFERENCES


CHEN, B. and PAN, G. M. (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when \( p/n \) both tend to infinity with their ratio converging to zero. Bernoulli 18 1405–1420. MR2995802 https://doi.org/10.3150/11-BEJ381


MSC2020 subject classifications. Primary 62H10; secondary 62H15.

Key words and phrases. Asymptotic normality, linear spectral statistics, general sample covariance matrix, ultra-dimension, matrix white noise, separable covariance.


Exponential Families in Theory and Practice

Bradley Efron, Stanford University

During the past half-century, exponential families have attained a position at the center of parametric statistical inference. Theoretical advances have been matched, and more than matched, in the world of applications, where logistic regression by itself has become the go-to methodology in medical statistics, computer-based prediction algorithms, and the social sciences. This book is based on a one-semester graduate course for first year Ph.D. and advanced master's students. After presenting the basic structure of univariate and multivariate exponential families, their application to generalized linear models including logistic and Poisson regression is described in detail, emphasizing geometrical ideas, computational practice, and the analogy with ordinary linear regression. Connections are made with a variety of current statistical methodologies: missing data, survival analysis and proportional hazards, false discovery rates, bootstrapping, and empirical Bayes analysis. The book connects exponential family theory with its applications in a way that doesn’t require advanced mathematical preparation.

www.imstat.org/cup/
The Institute of Mathematical Statistics presents

**IMS MONOGRAPH**

**Case-Control Studies**

Ruth H. Keogh
and D. R. Cox

The case-control approach is a powerful method for investigating factors that may explain a particular event. It is extensively used in epidemiology to study disease incidence, one of the best-known examples being Bradford Hill and Doll’s investigation of the possible connection between cigarette smoking and lung cancer. More recently, case-control studies have been increasingly used in other fields, including sociology and econometrics.

With a particular focus on statistical analysis, this book is ideal for applied and theoretical statisticians wanting an up-to-date introduction to the field. It covers the fundamentals of case-control study design and analysis as well as more recent developments, including two-stage studies, case-only studies and methods for case-control sampling in time. The latter have important applications in large prospective cohorts which require case-control sampling designs to make efficient use of resources. More theoretical background is provided in an appendix for those new to the field.