Vol. 48, No. 3—June 2020

THE ANNALS of
STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Testing in high-dimensional spiked models ............... IAIN M. JOHNSTONE AND ALEXEI ONATSKI 1231

Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices .... T. TONY CAI, XIAO HAN AND GUANGMING PAN 1255

Post hoc confidence bounds on false positives using reference families GILLES BLANCHARD, PIERRE NEUVIAL AND ETIENNE ROQUAIN 1281

Distribution and correlation-free two-sample test of high-dimensional means KAIJIE XUE AND FANG YAO 1304

Just interpolate: Kernel “Ridgeless” regression can generalize TENGYUAN LIANG AND ALEXANDER RAKHLIN 1329

Bridging the gap between constant step size stochastic gradient descent and Markov chains ........ AYMERIC DIEULEVEUT, ALAIN DURMU S AND FRANCIS BACH 1348

Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions RICHARD NICKL AND KOLYAN RAY 1383

Robust inference with knockoffs RINA FOYGEL BARBER, EMANUEL J. CANDÈS AND RICHARD J. SAMWORTH 1409

Nonparametric Bayesian analysis of the compound Poisson prior for support boundary recovery ............ MARKUS REISS AND JOHANNES SCHMIDT-HIEBER 1432

Entrywise eigenvector analysis of random matrices with low expected rank EMANUEL ABBE, JIANQING FAN, KAIZHENG WANG AND YIQIAO ZHONG 1452

Concentration of tempered posteriors and of their variational approximations PIERRE ALQUIER AND JAMES RIDGWAY 1475

Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image NICHOLAS SYRING AND RYAN MARTIN 1498

The hardness of conditional independence testing and the generalised covariance measure RAJEN D. SHAH AND JONAS PETERS 1514

Some theoretical properties of GANS GÉRARD BIAU, BENOÎT CADRE, MAXIME SANGNIER AND UGO TANIELIAN 1539

On post dimension reduction statistical inference KYONGWON KIM, BING LI, ZHOU YU AND LEXIN LI 1567

Statistical and computational limits for sparse matrix detection T. TONY CAI AND YIHONG WU 1593

Segmentation and estimation of change-point models: False positive control and confidence regions ............ XIAO FANG, JIAN LI AND DAVID SIEGMUND 1615

Robust covariance estimation under \( L_4 - L_2 \) norm equivalence SHAHAR MENDELSON AND NIKITA ZHIVOTOVSKIY 1648

Robust inference via multiplier bootstrap ............... XI CHEN AND WEN-XIN ZHOU 1665

On the optimal reconstruction of partially observed functional data ALOIS KNEIP AND DOMINIK LIEBL 1692

Large sample properties of partitioning-based series estimators MATIAS D. CAT TANE, MAX H. FARRELL AND YINGJIE FENG 1718

Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score ....... QINGYUAN ZHAO, JINGSHU WANG, GIBRAN HEMANI, J ACK BOWDEN AND DYLAN S. SMALL 1742

continued
Local uncertainty sampling for large-scale multiclass logistic regression
LEI HAN, KEAN MING TAN, TING YANG AND TONG ZHANG 1770

Local nearest neighbour classification with applications to semi-supervised learning
TIMOTHY I. CANNINGS, THOMAS B. BERRETT AND RICHARD J. SAMWORTH 1789

An adaptable generalization of Hotelling’s $T^2$ test in high dimension
HAORAN LI, ALEXANDER AUE, DEBASHIS PAUL, JIE PENG AND PEI WANG 1815

GRID: A variable selection and structure discovery method for high dimensional nonparametric regression
FRANCESCO GIORDANO, SOUMENDRA NATH LAHIRI AND MARIA LUCIA PARRELLA 1848
TESTING IN HIGH-DIMENSIONAL SPIKED MODELS

BY IAIN M. JOHNSTONE\textsuperscript{1} AND ALEXEI ONATS|KI\textsuperscript{2}

\textsuperscript{1}Department of Statistics, Stanford University, imj@stanford.edu
\textsuperscript{2}Faculty of Economics, University of Cambridge, ao319@cam.ac.uk

We consider the five classes of multivariate statistical problems identified by James (Ann. Math. Stat. 35 (1964) 475–501), which together cover much of classical multivariate analysis, plus a simpler limiting case, symmetric matrix denoising. Each of James’ problems involves the eigenvalues of $E^{-1}H$ where $H$ and $E$ are proportional to high-dimensional Wishart matrices. Under the null hypothesis, both Wisharts are central with identity covariance. Under the alternative, the noncentrality or the covariance parameter of $H$ has a single eigenvalue, a spike, that stands alone. When the spike is smaller than a case-specific phase transition threshold, none of the sample eigenvalues separate from the bulk, making the testing problem challenging. Using a unified strategy for the six cases, we show that the log likelihood ratio processes parameterized by the value of the subcritical spike converge to Gaussian processes with logarithmic correlation. We then derive asymptotic power envelopes for tests for the presence of a spike.

REFERENCES


MSC2010 subject classifications. Primary 62E20; secondary 62H15.

Key words and phrases. Likelihood ratio test, hypergeometric function, principal components analysis, canonical correlations, matrix denoising, multiple response regression.
LIMITING LAWS FOR DIVERGENT SPIKED EIGENVALUES AND LARGEST NONSPIKED EIGENVALUE OF SAMPLE COVARIANCE MATRICES

BY T. TONY CAI1, XIAO HAN2 AND GUANGMING PAN3

1Department of Statistics, The Wharton School, University of Pennsylvania, tcai@wharton.upenn.edu
2University of Science and Technology of China and Nanyang Technological University, xhan011@e.ntu.edu.sg
3School of Physical & Mathematical Sciences, Nanyang Technological University, gmpan@ntu.edu.sg

We study the asymptotic distributions of the spiked eigenvalues and the largest nonspiked eigenvalue of the sample covariance matrix under a general covariance model with divergent spiked eigenvalues, while the other eigenvalues are bounded but otherwise arbitrary. The limiting normal distribution for the spiked sample eigenvalues is established. It has distinct features that the asymptotic mean relies on not only the population spikes but also the nonspikes and that the asymptotic variance in general depends on the population eigenvectors. In addition, the limiting Tracy–Widom law for the largest nonspiked sample eigenvalue is obtained.

Estimation of the number of spikes and the convergence of the leading eigenvectors are also considered. The results hold even when the number of the spikes diverges. As a key technical tool, we develop a central limit theorem for a type of random quadratic forms where the random vectors and random matrices involved are dependent. This result can be of independent interest.

REFERENCES


MSC2010 subject classifications. Primary 62H25, 60B20; secondary 60F05, 62H10.

Key words and phrases. Extreme eigenvalues, factor model, principal component analysis, sample covariance matrix, spiked covariance matrix model, Tracy–Widom distribution.


POST HOC CONFIDENCE BOUNDS ON FALSE POSITIVES USING REFERENCE FAMILIES

BY GILLES BLANCHARD¹, PIERRE NEUVIAL² AND ETIENNE ROQUAIN³

¹Institut für Mathematik, Universität Potsdam, gilles.blanchard@math.uni-potsdam.de
²Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, pierre.neuvial@math.univ-toulouse.fr
³Sorbonne Université, Sorbonne Paris Cité, CNRS, Laboratoire de Probabilités Statistique et Modélisation, LPSM, etienne.roquain@upmc.fr

We follow a post hoc, “user-agnostic” approach to false discovery control in a large-scale multiple testing framework, as introduced by Genovese and Wasserman [J. Amer. Statist. Assoc. 101 (2006) 1408–1417], Goeman and Solari [Statist. Sci. 26 (2011) 584–597]: the statistical guarantee on the number of correct rejections must hold for any set of candidate items, possibly selected by the user after having seen the data. To this end, we introduce a novel point of view based on a family of reference rejection sets and a suitable criterion, namely the joint familywise error rate over that family (JER for short). First, we establish how to derive post hoc bounds from a given JER control and analyze some general properties of this approach. We then develop procedures for controlling the JER in the case where reference regions are \(p\)-value level sets. These procedures adapt to dependencies and to the unknown quantity of signal (via a step-down principle). We also show interesting connections to confidence envelopes of Meinshausen [Scand. J. Stat. 33 (2006) 227–237]; Genovese and Wasserman [J. Amer. Statist. Assoc. 101 (2006) 1408–1417], the closed testing based approach of Goeman and Solari [Statist. Sci. 26 (2011) 584–597] and to the higher criticism of Donoho and Jin [Ann. Statist. 32 (2004) 962–994]. Our theoretical statements are supported by numerical experiments.

REFERENCES


MSC2010 subject classifications. Primary 62G10; secondary 62H15.

Key words and phrases. Post hoc inference, multiple testing, Simes inequality, family-wise error rate, step-down algorithm, dependence, higher criticism.
We propose a two-sample test for high-dimensional means that requires neither distributional nor correlational assumptions, besides some weak conditions on the moments and tail properties of the elements in the random vectors. This two-sample test based on a nontrivial extension of the one-sample central limit theorem (Ann. Probab. 45 (2017) 2309–2352) provides a practically useful procedure with rigorous theoretical guarantees on its size and power assessment. In particular, the proposed test is easy to compute and does not require the independently and identically distributed assumption, which is allowed to have different distributions and arbitrary correlation structures. Further desired features include weaker moments and tail conditions than existing methods, allowance for highly unequal sample sizes, consistent power behavior under fairly general alternative, data dimension allowed to be exponentially high under the umbrella of such general conditions. Simulated and real data examples have demonstrated favorable numerical performance over existing methods.

REFERENCES


MSC2010 subject classifications. 62H05, 62F05.
Key words and phrases. High-dimensional central limit theorem, Kolmogorov distance, multiplier bootstrap, power function.


JUST INTERPOLATE: KERNEL “RIDGELESS” REGRESSION CAN GENERALIZE

BY TENGYUAN LIANG\(^1\) AND ALEXANDER RAKHLIN\(^2\)

\(^1\)Econometrics and Statistics, Booth School of Business, University of Chicago, tengyuan.liang@chicagobooth.edu
\(^2\)Center for Statistics & IDSS, Massachusetts Institute of Technology, rakhlin@mit.edu

In the absence of explicit regularization, Kernel “Ridgeless” Regression with nonlinear kernels has the potential to fit the training data perfectly. It has been observed empirically, however, that such interpolated solutions can still generalize well on test data. We isolate a phenomenon of implicit regularization for minimum-norm interpolated solutions which is due to a combination of high dimensionality of the input data, curvature of the kernel function and favorable geometric properties of the data such as an eigenvalue decay of the empirical covariance and kernel matrices. In addition to deriving a data-dependent upper bound on the out-of-sample error, we present experimental evidence suggesting that the phenomenon occurs in the MNIST dataset.

REFERENCES


MSC2010 subject classifications. 68Q32, 62G08.
Key words and phrases. Minimum-norm interpolation, reproducing kernel Hilbert spaces, implicit regularization, high dimensionality, data-dependent bounds, kernel methods, spectral decay.
BRIDGING THE GAP BETWEEN CONSTANT STEP SIZE STOCHASTIC GRADIENT DESCENT AND MARKOV CHAINS

BY AYMERIC DIEULEVEUT1, ALAIN DURMUSS AND FRANCIS BACH3

1CMAP, École Polytechnique (X), aymeric.dieuleveut@polytechnique.edu
2CMLA - École normale supérieure Paris-Saclay, CNRS, Université Paris-Saclay, alain.durmus@cmla.ens-cachan.fr
3INRIA - Département d’informatique de l’ENS, École normale supérieure, CNRS, PSL Research University, francis.bach@inria.fr

We consider the minimization of a strongly convex objective function given access to unbiased estimates of its gradient through stochastic gradient descent (SGD) with constant step size. While the detailed analysis was only performed for quadratic functions, we provide an explicit asymptotic expansion of the moments of the averaged SGD iterates that outlines the dependence on initial conditions, the effect of noise and the step size, as well as the lack of convergence in the general (nonquadratic) case. For this analysis we bring tools from Markov chain theory into the analysis of stochastic gradient. We then show that Richardson–Romberg extrapolation may be used to get closer to the global optimum, and we show empirical improvements of the new extrapolation scheme.

REFERENCES


MSC2010 subject classifications. Primary 62L20; secondary 90C15, 93E35.

Key words and phrases. Stochastic gradient descent, Markov chains.


NONPARAMETRIC STATISTICAL INFERENCE FOR DRIFT VECTOR FIELDS OF MULTI-DIMENSIONAL DIFFUSIONS

BY RICHARD NICKL$^1$ AND KOLYAN RAY$^2$

$^1$Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, r.nickl@statslab.cam.ac.uk

$^2$Department of Mathematics, King’s College London, kolyan.ray@kcl.ac.uk

The problem of determining a periodic Lipschitz vector field $b = (b_1, \ldots, b_d)$ from an observed trajectory of the solution $(X_t : 0 \leq t \leq T)$ of the multi-dimensional stochastic differential equation

$$dX_t = b(X_t) \, dt + dW_t, \quad t \geq 0,$$

where $W_t$ is a standard $d$-dimensional Brownian motion, is considered. Convergence rates of a penalised least squares estimator, which equals the maximum a posteriori (MAP) estimate corresponding to a high-dimensional Gaussian product prior, are derived. These results are deduced from corresponding contraction rates for the associated posterior distributions. The rates obtained are optimal up to log-factors in $L^2$-loss in any dimension, and also for supremum norm loss when $d \leq 4$. Further, when $d \leq 3$, nonparametric Bernstein–von Mises theorems are proved for the posterior distributions of $b$. From this, we deduce functional central limit theorems for the implied estimators of the invariant measure $\mu_b$. The limiting Gaussian process distributions have a covariance structure that is asymptotically optimal from an information-theoretic point of view.

REFERENCES


Key words and phrases. Penalised least squares estimator, asymptotics of nonparametric Bayes procedures, Bernstein–von Mises theorem, uncertainty quantification.
We consider the variable selection problem, which seeks to identify important variables influencing a response $Y$ out of many candidate features $X_1,\ldots, X_p$. We wish to do so while offering finite-sample guarantees about the fraction of false positives—selected variables $X_j$ that in fact have no effect on $Y$ after the other features are known. When the number of features $p$ is large (perhaps even larger than the sample size $n$), and we have no prior knowledge regarding the type of dependence between $Y$ and $X$, the model-X knockoffs framework nonetheless allows us to select a model with a guaranteed bound on the false discovery rate, as long as the distribution of the feature vector $X = (X_1,\ldots, X_p)$ is exactly known. This model selection procedure operates by constructing “knockoff copies” of each of the $p$ features, which are then used as a control group to ensure that the model selection algorithm is not choosing too many irrelevant features. In this work, we study the practical setting where the distribution of $X$ can only be estimated, rather than known exactly, and the knockoff copies of the $X_j$’s are therefore constructed somewhat incorrectly. Our results, which are free of any modeling assumption whatsoever, show that the resulting model selection procedure incurs an inflation of the false discovery rate that is proportional to our errors in estimating the distribution of each feature $X_j$ conditional on the remaining features $\{X_k : k \neq j\}$. The model-X knockoffs framework is therefore robust to errors in the underlying assumptions on the distribution of $X$, making it an effective method for many practical applications, such as genome-wide association studies, where the underlying distribution on the features $X_1,\ldots, X_p$ is estimated accurately but not known exactly.

REFERENCES


Key words and phrases. Knockoffs, variable selection, false discovery rate (FDR), high-dimensional regression, robustness.


NONPARAMETRIC BAYESIAN ANALYSIS OF THE COMPOUND POISSON PRIOR FOR SUPPORT BOUNDARY RECOVERY

BY MARKUS REISS\(^1\) AND JOHANNES SCHMIDT-HIEBER\(^2\)

\(^1\)Institut für Mathematik, Humboldt-Universität zu Berlin, mreiss@math.hu-berlin.de
\(^2\)Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl

Given data from a Poisson point process with intensity \( (x, y) \mapsto n1(f(x) \leq y) \), frequentist properties for the Bayesian reconstruction of the support boundary function \( f \) are derived. We mainly study compound Poisson process priors with fixed intensity proving that the posterior contracts with nearly optimal rate for monotone support boundaries and adapts to Hölder smooth boundaries. We then derive a limiting shape result for a compound Poisson process prior and a function space with increasing parameter dimension. It is shown that the marginal posterior of the mean functional performs an automatic bias correction and contracts with a faster rate than the MLE. In this case, \((1 - \alpha)\)-credible sets are also asymptotic \((1 - \alpha)\)-confidence intervals. As a negative result, it is shown that the frequentist coverage of credible sets is lost for linear functions \( f \) outside the function class.

REFERENCES


MSC2010 subject classifications. Primary 62C10, 62G05; secondary 60G55.

Key words and phrases. Frequentist Bayes analysis, posterior contraction, Bernstein–von Mises theorem, Poisson point process, boundary detection, compound Poisson process, subordinator prior.
ENTRYWISE EIGENVECTOR ANALYSIS OF RANDOM MATRICES WITH LOW EXPECTED RANK

BY EMMANUEL ABBE\textsuperscript{1}, JIANQING FAN\textsuperscript{2,*}, KAIZHENG WANG\textsuperscript{2,**} AND YIQIAO ZHONG\textsuperscript{2,†}

\textsuperscript{1}PACM and Department of EE, Princeton University, eabbe@princeton.edu
\textsuperscript{2}Department of ORFE, Princeton University, \textsuperscript{*}jqfan@princeton.edu; \textsuperscript{**}kaizheng@princeton.edu; \textsuperscript{†}yiqiao@princeton.edu

Recovering low-rank structures via eigenvector perturbation analysis is a common problem in statistical machine learning, such as in factor analysis, community detection, ranking, matrix completion, among others. While a large variety of bounds are available for average errors between empirical and population statistics of eigenvectors, few results are tight for entrywise analyses, which are critical for a number of problems such as community detection.

This paper investigates entrywise behaviors of eigenvectors for a large class of random matrices whose expectations are low rank, which helps settle the conjecture in Abbe, Bandeira and Hall (2014) that the spectral algorithm achieves exact recovery in the stochastic block model without any trimming or cleaning steps. The key is a first-order approximation of eigenvectors under the $\ell_\infty$ norm:

$$u_k \approx \frac{Au_k}{\lambda_k},$$

where \{u_k\} and \{u_k^\star\} are eigenvectors of a random matrix $A$ and its expectation $E A$, respectively. The fact that the approximation is both tight and linear in $A$ facilitates sharp comparisons between $u_k$ and $u_k^\star$. In particular, it allows for comparing the signs of $u_k$ and $u_k^\star$ even if $\|u_k - u_k^\star\|_\infty$ is large. The results are further extended to perturbations of eigenspaces, yielding new $\ell_\infty$-type bounds for synchronization ($\mathbb{Z}_2$-spiked Wigner model) and noisy matrix completion.

REFERENCES


\textbf{MSC2010 subject classifications.} Primary 62H25; secondary 60B20, 62H12.

\textbf{Key words and phrases.} Eigenvector perturbation, spectral analysis, synchronization, community detection, matrix completion, low-rank structures, random matrices.


CONCENTRATION OF TEMPERED POSTERIORS AND OF THEIR VARIATIONAL APPROXIMATIONS

BY PIERRE ALQUIER\textsuperscript{1} AND JAMES RIDGWAY\textsuperscript{2}

\textsuperscript{1}RIKEN Center for Advanced Intelligence Project, pierre.alquier.stat@gmail.com
\textsuperscript{2}Capital Fund Management, james.lp.ridgway@gmail.com

While Bayesian methods are extremely popular in statistics and machine learning, their application to massive data sets is often challenging, when possible at all. The classical MCMC algorithms are prohibitively slow when both the model dimension and the sample size are large. Variational Bayesian methods aim at approximating the posterior by a distribution in a tractable family $\mathcal{F}$. Thus, MCMC are replaced by an optimization algorithm which is orders of magnitude faster. VB methods have been applied in such computationally demanding applications as collaborative filtering, image and video processing or NLP to name a few. However, despite nice results in practice, the theoretical properties of these approximations are not known. We propose a general oracle inequality that relates the quality of the VB approximation to the prior $\pi$ and to the structure of $\mathcal{F}$. We provide a simple condition that allows to derive rates of convergence from this oracle inequality. We apply our theory to various examples. First, we show that for parametric models with log-Lipschitz likelihood, Gaussian VB leads to efficient algorithms and consistent estimators. We then study a high-dimensional example: matrix completion, and a nonparametric example: density estimation.

REFERENCES


Key words and phrases. Concentration of the posterior, rate of convergence, variational approximation, PAC-Bayesian bounds.


ROBUST AND RATE-OPTIMAL GIBBS POSTERIOR INFERENCE ON THEBOUNDARY OF A NOISY IMAGE

BY NICHOLAS SYRING\(^1\) AND RYAN MARTIN\(^2\)

\(^1\)Department of Mathematics and Statistics, Washington University in St. Louis, nasyring@wustl.edu
\(^2\)Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

Detection of an image boundary when the pixel intensities are measured with noise is an important problem in image segmentation. From a statistical point of view, a challenge is that likelihood-based methods require modeling the pixel intensities inside and outside the image boundary, even though these distributions are typically not of interest. Since misspecification of the pixel intensity distributions can negatively affect inference on the image boundary, it would be desirable to avoid this modeling step altogether. Toward this, we develop a robust Gibbsian approach that constructs a posterior distribution for the image boundary directly, without modeling the pixel intensities. We prove that the Gibbs posterior concentrates asymptotically at the minimax optimal rate, adaptive to the boundary smoothness. Monte Carlo computation of the Gibbs posterior is straightforward, and simulation results show that the corresponding inference is more accurate than that based on existing Bayesian methodology.

REFERENCES


MSC2010 subject classifications. Primary 62G20; secondary 62G05.

Key words and phrases. Adaptation, boundary detection, likelihood-free inference, model misspecification, posterior concentration rate.


THE HARDNESS OF CONDITIONAL INDEPENDENCE TESTING AND
THE GENERALISED COVARIANCE MEASURE

BY RAJEN D. SHAH\(^1\) AND JONAS PETERS\(^2\)

\(^1\)Statistical Laboratory, University of Cambridge, r.shah@statslab.cam.ac.uk
\(^2\)Department of Mathematical Sciences, University of Copenhagen, jonas.peters@math.ku.dk

It is a common saying that testing for conditional independence, that is, testing whether whether two random vectors \(X\) and \(Y\) are independent, given \(Z\), is a hard statistical problem if \(Z\) is a continuous random variable (or vector). In this paper, we prove that conditional independence is indeed a particularly difficult hypothesis to test for. Valid statistical tests are required to have a size that is smaller than a pre-defined significance level, and different tests usually have power against a different class of alternatives. We prove that a valid test for conditional independence does not have power against any alternative.

Given the nonexistence of a uniformly valid conditional independence test, we argue that tests must be designed so their suitability for a particular problem may be judged easily. To address this need, we propose in the case where \(X\) and \(Y\) are univariate to nonlinearly regress \(X\) on \(Z\), and \(Y\) on \(Z\) and then compute a test statistic based on the sample covariance between the residuals, which we call the generalised covariance measure (GCM). We prove that validity of this form of test relies almost entirely on the weak requirement that the regression procedures are able to estimate the conditional means \(X\) given \(Z\), and \(Y\) given \(Z\), as low rates. We extend the methodology to handle settings where \(X\) and \(Y\) may be multivariate or even high dimensional. While our general procedure can be tailored to the setting at hand by combining it with any regression technique, we develop the theoretical guarantees for kernel ridge regression. A simulation study shows that the test based on GCM is competitive with state of the art conditional independence tests. Code is available as the R package GeneralisedCovarianceMeasure on CRAN.

REFERENCES


**MSC2010 subject classifications.** Primary 62G10; secondary 62G08.

**Key words and phrases.** Conditional independence, hypothesis testing, testability, wild bootstrap, kernel ridge regression.


SOME THEORETICAL PROPERTIES OF GANS

BY GÉRARD BIAU1,*, BENOÎT CADRE2, MAXIME SANGNIER1,** AND UGO TANIELIAN1,†

1Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, *gerard.biau@sorbonne-universite.fr; **maxime.sangnier@sorbonne-universite.fr; †u.tanielian@criteo.com
2ENS Rennes, Univ Rennes, IRMAR, benoit.cadre@ens-rennes.fr

Generative Adversarial Networks (GANs) are a class of generative algorithms that have been shown to produce state-of-the-art samples, especially in the domain of image creation. The fundamental principle of GANs is to approximate the unknown distribution of a given data set by optimizing an objective function through an adversarial game between a family of generators and a family of discriminators. In this paper, we offer a better theoretical understanding of GANs by analyzing some of their mathematical and statistical properties. We study the deep connection between the adversarial principle underlying GANs and the Jensen–Shannon divergence, together with some optimality characteristics of the problem. An analysis of the role of the discriminator family via approximation arguments is also provided. In addition, taking a statistical point of view, we study the large sample properties of the estimated distribution and prove in particular a central limit theorem. Some of our results are illustrated with simulated examples.

REFERENCES


MSC2010 subject classifications. Primary 62F12; secondary 68T01.
Key words and phrases. Generative models, adversarial principle, Jensen–Shannon divergence, neural networks, central limit theorem.


ON POST DIMENSION REDUCTION STATISTICAL INFERENCE

BY KYONGWON KIM1,* , BING LI1,**, ZHOU YU2 AND LEXIN LI3

1Department of Statistics, Pennsylvania State University; *kbk5206@psu.edu; **bxl9@psu.edu
2KLATASDS-MOE, School of Statistics, East China Normal University, zyu@stat.ecnu.edu.cn
3Department of Biostatistics and Epidemiology, University of California, Berkeley, lexinli@berkeley.edu

The methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades. However, there has been a lack of systematic and rigorous development of post dimension reduction inference, which has seriously hindered its applications. The current common practice is to treat the estimated sufficient predictors as the true predictors and use them as the starting point of the downstream statistical inference. However, this naive inference approach would grossly overestimate the confidence level of an interval, or the power of a test, leading to the distorted results. In this paper, we develop a general and comprehensive framework of post dimension reduction inference, which can accommodate any dimension reduction method and model building method, as long as their corresponding influence functions are available. Within this general framework, we derive the influence functions and present the explicit post reduction formulas for the combinations of numerous dimension reduction and model building methods. We then develop post reduction inference methods for both confidence interval and hypothesis testing. We investigate the finite-sample performance of our procedures by simulations and a real data analysis.

REFERENCES


MSC2010 subject classifications. Primary 62G08; secondary 62H99.

Key words and phrases. Central subspace, directional regression, estimating equations, generalized method of moment, influence function, sliced inverse regression, Von Mises expansion.


STATIONARY AND COMPUTATIONAL LIMITS FOR SPARSE MATRIX DETECTION

BY T. TONY CAI1 AND YIHONG WU2

1 Department of Statistics, The Wharton School, University of Pennsylvania, tcai@wharton.upenn.edu
2 Department of Statistics and Data Science, Yale University, yihong.wu@yale.edu

This paper investigates the fundamental limits for detecting a high-dimensional sparse matrix contaminated by white Gaussian noise from both the statistical and computational perspectives. We consider $p \times p$ matrices whose rows and columns are individually $k$-sparse. We provide a tight characterization of the statistical and computational limits for sparse matrix detection, which precisely describe when achieving optimal detection is easy, hard or impossible, respectively. Although the sparse matrices considered in this paper have no apparent submatrix structure and the corresponding estimation problem has no computational issue at all, the detection problem has a surprising computational barrier when the sparsity level $k$ exceeds the cubic root of the matrix size $p$: attaining the optimal detection boundary is computationally at least as hard as solving the planted clique problem.

The same statistical and computational limits also hold in the sparse covariance matrix model, where each variable is correlated with at most $k$ others. A key step in the construction of the statistically optimal test is a structural property of sparse matrices, which can be of independent interest.

REFERENCES


Key words and phrases. Minimax rates, computational limits, sparse covariance matrix, sparse detection.
REFERENCES


SEGMENTATION AND ESTIMATION OF CHANGE-POINT MODELS: FALSE POSITIVE CONTROL AND CONFIDENCE REGIONS

BY XIAO FANG\textsuperscript{1}, JIAN LI\textsuperscript{2} AND DAVID SIEGMUND\textsuperscript{3}

\textsuperscript{1}Department of Statistics, The Chinese University of Hong Kong, xfang@sta.cuhk.edu.hk
\textsuperscript{2}Adobe Systems, jianli0124@gmail.com
\textsuperscript{3}Department of Statistics, Stanford University, siegmund@stanford.edu

To segment a sequence of independent random variables at an unknown number of change-points, we introduce new procedures that are based on thresholding the likelihood ratio statistic, and give approximations for the probability of a false positive error when there are no change-points. We also study confidence regions based on the likelihood ratio statistic for the change-points and joint confidence regions for the change-points and the parameter values. Applications to segment array CGH data are discussed.

REFERENCES


MSC2010 subject classifications. 62G05, 62G15.  
Key words and phrases. Array CGH analysis, change-points, confidence regions, exponential families, likelihood ratio statistics.


WORSLEY, K. J. (1986). Confidence regions and test for a change-point in a sequence of exponential family random variables. *Biometrika* 73 91–104. MR0836437 https://doi.org/10.1093/biomet/73.1.91


ROBUST COVARIANCE ESTIMATION UNDER \( L_4 - L_2 \) NORM EQUIVALENCE

BY SHAHAR MENDELSON\(^1\) AND NIKITA ZHIVOTOVSKIY\(^2\)

\(^1\)Mathematical Sciences Institute, Australian National University, shahar.mendelson@upmc.fr

\(^2\)Higher School of Economics, nikita.zhivotovskiy@phystech.edu

Let \( X \) be a centered random vector taking values in \( \mathbb{R}^d \) and let \( \Sigma = \mathbb{E}(X \otimes X) \) be its covariance matrix. We show that if \( X \) satisfies an \( L_4 - L_2 \) norm equivalence (sometimes referred to as the bounded kurtosis assumption), there is a covariance estimator \( \hat{\Sigma} \) that exhibits almost the same performance one would expect had \( X \) been a Gaussian vector. The procedure also improves the current state-of-the-art regarding high probability bounds in the sub-Gaussian case (sharp results were only known in expectation or with constant probability).

In both scenarios the new bounds do not depend explicitly on the dimension \( d \), but rather on the effective rank of the covariance matrix \( \Sigma \).

REFERENCES


MSC2010 subject classifications. Primary 62G35; secondary 62G15.

Key words and phrases. Covariance estimation, robust estimation, median of means.

This paper investigates the theoretical underpinnings of two fundamental statistical inference problems, the construction of confidence sets and large-scale simultaneous hypothesis testing, in the presence of heavy-tailed data. With heavy-tailed observation noise, finite sample properties of the least squares-based methods, typified by the sample mean, are suboptimal both theoretically and empirically. In this paper, we demonstrate that the adaptive Huber regression, integrated with the multiplier bootstrap procedure, provides a useful robust alternative to the method of least squares. Our theoretical and empirical results reveal the effectiveness of the proposed method, and highlight the importance of having inference methods that are robust to heavy tailedness.

REFERENCES


Key words and phrases. Confidence set, heavy-tailed data, multiple testing, multiplier bootstrap, robust regression, Wilks’ theorem.


https://doi.org/10.1214/17-AOS1606
ON THE OPTIMAL RECONSTRUCTION OF PARTIALLY OBSERVED FUNCTIONAL DATA

BY ALOIS KNEIP * AND DOMINIK LIEBL **

Statistische Abteilung, University of Bonn, *akneip@uni-bonn.de; **dliebl@uni-bonn.de

We propose a new reconstruction operator that aims to recover the missing parts of a function given the observed parts. This new operator belongs to a new, very large class of functional operators which includes the classical regression operators as a special case. We show the optimality of our reconstruction operator and demonstrate that the usually considered regression operators generally cannot be optimal reconstruction operators. Our estimation theory allows for autocorrelated functional data and considers the practically relevant situation in which each of the $n$ functions is observed at $m_i$, $i = 1, \ldots, n$, discretization points. We derive rates of consistency for our nonparametric estimation procedures using a double asymptotic. For data situations, as in our real data application where $m_i$ is considerably smaller than $n$, we show that our functional principal components based estimator can provide better rates of convergence than conventional nonparametric smoothing methods.

REFERENCES


Key words and phrases. Functional data analysis, functional principal components, incomplete functions.


LARGE SAMPLE PROPERTIES OF PARTITIONING-BASED SERIES ESTIMATORS

BY MATIAS D. CATTANEO¹, MAX H. FARRELL² AND YINGJIE FENG³

¹Department of Operations Research and Financial Engineering, Princeton University, cattaneo@princeton.edu
²Booth School of Business, University of Chicago, max.farrell@chicagobooth.edu
³Department of Politics, Princeton University, yingjief@princeton.edu

We present large sample results for partitioning-based least squares nonparametric regression, a popular method for approximating conditional expectation functions in statistics, econometrics and machine learning. First, we obtain a general characterization of their leading asymptotic bias. Second, we establish integrated mean squared error approximations for the point estimator and propose feasible tuning parameter selection. Third, we develop pointwise inference methods based on undersmoothing and robust bias correction. Fourth, employing different coupling approaches, we develop uniform distributional approximations for the undersmoothed and robust bias-corrected t-statistic processes and construct valid confidence bands. In the univariate case, our uniform distributional approximations require seemingly minimal rate restrictions and improve on approximation rates known in the literature. Finally, we apply our general results to three partitioning-based estimators: splines, wavelets and piecewise polynomials. The Supplemental Appendix includes several other general and example-specific technical and methodological results. A companion R package is provided.

REFERENCES


Key words and phrases. Nonparametric regression, series methods, sieve methods, robust bias correction, uniform inference, strong approximation, tuning parameter selection.


Mendelian randomization (MR) is a method of exploiting genetic variation to unbiasedly estimate a causal effect in presence of unmeasured confounding. MR is being widely used in epidemiology and other related areas of population science. In this paper, we study statistical inference in the increasingly popular two-sample summary-data MR design. We show a linear model for the observed associations approximately holds in a wide variety of settings when all the genetic variants satisfy the exclusion restriction assumption, or in genetic terms, when there is no pleiotropy. In this scenario, we derive a maximum profile likelihood estimator with provable consistency and asymptotic normality. However, through analyzing real datasets, we find strong evidence of both systematic and idiosyncratic pleiotropy in MR, echoing the omnigenic model of complex traits that is recently proposed in genetics. We model the systematic pleiotropy by a random effects model, where no genetic variant satisfies the exclusion restriction condition exactly. In this case, we propose a consistent and asymptotically normal estimator by adjusting the profile score. We then tackle the idiosyncratic pleiotropy by robustifying the adjusted profile score. We demonstrate the robustness and efficiency of the proposed methods using several simulated and real datasets.

REFERENCES


MSC2010 subject classifications. Primary 65J05; secondary 46N60, 62F35.

Key words and phrases. Causal inference, limited information maximum likelihood, weak instruments, errors in variables, path analysis, pleiotropy effects.
LOCAL UNCERTAINTY SAMPLING FOR LARGE-SCALE MULTICLASS LOGISTIC REGRESSION

BY LEI HAN¹, KEAN MING TAN², TING YANG³ AND TONG ZHANG⁴

¹Tencent AI Lab, Tencent Technology (Shenzhen) Company Limited, leihan.cs@gmail.com
²Department of Statistics, University of Michigan, keanming@umich.edu
³Yelp Inc., ting.yang.pku@gmail.com
⁴Department of Mathematics and Department of Computer Science, Hong Kong University of Science and Technology, tongzhang@tongzhang-ml.org

A major challenge for building statistical models in the big data era is that the available data volume far exceeds the computational capability. A common approach for solving this problem is to employ a subsampled dataset that can be handled by available computational resources. We propose a general subsampling scheme for large-scale multiclass logistic regression and examine the variance of the resulting estimator. We show that asymptotically, the proposed method always achieves a smaller variance than that of the uniform random sampling. Moreover, when the classes are conditionally imbalanced, significant improvement over uniform sampling can be achieved. Empirical performance of the proposed method is evaluated and compared to other methods via both simulated and real-world datasets, and these results match and confirm our theoretical analysis.

REFERENCES


MSC2010 subject classifications. Primary 62D05; secondary 62F10, 62F12.

Key words and phrases. Sampling, large-scale, multiclass logistic regression.


LOCAL NEAREST NEIGHBOUR CLASSIFICATION WITH APPLICATIONS TO SEMI-SUPERVISED LEARNING

BY TIMOTHY I. CANNINGS¹, THOMAS B. BERRETT²,* AND RICHARD J. SAMWORTH²,**

¹School of Mathematics, University of Edinburgh, timothy.cannings@ed.ac.uk
²Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, *t.berrett@statslab.cam.ac.uk; **r.samworth@statslab.cam.ac.uk

We derive a new asymptotic expansion for the global excess risk of a local-$k$-nearest neighbour classifier, where the choice of $k$ may depend upon the test point. This expansion elucidates conditions under which the dominant contribution to the excess risk comes from the decision boundary of the optimal Bayes classifier, but we also show that if these conditions are not satisfied, then the dominant contribution may arise from the tails of the marginal distribution of the features. Moreover, we prove that, provided the $d$-dimensional marginal distribution of the features has a finite $\rho$th moment for some $\rho > 4$ (as well as other regularity conditions), a local choice of $k$ can yield a rate of convergence of the excess risk of $O(n^{-4/(d+4)})$, where $n$ is the sample size, whereas for the standard $k$-nearest neighbour classifier, our theory would require $d \geq 5$ and $\rho > 4d/(d-4)$ finite moments to achieve this rate. These results motivate a new $k$-nearest neighbour classifier for semi-supervised learning problems, where the unlabelled data are used to obtain an estimate of the marginal feature density, and fewer neighbours are used for classification when this density estimate is small. Our worst-case rates are complemented by a minimax lower bound, which reveals that the local, semi-supervised $k$-nearest neighbour classifier attains the minimax optimal rate over our classes for the excess risk, up to a subpolynomial factor in $n$. These theoretical improvements over the standard $k$-nearest neighbour classifier are also illustrated through a simulation study.

REFERENCES


MSC2010 subject classifications. 62G20.
Key words and phrases. Classification problems, nearest neighbours, nonparametric classification, semi-supervised learning.
AN ADAPTABLE GENERALIZATION OF HOTELLING’S $T^2$ TEST IN HIGH DIMENSION

BY HAORAN LI$^1$,*, ALEXANDER AUE$^1$,**, DEBASHIS PAUL$^1$,†, JIE PENG$^1$,‡ AND PEI WANG$^2$

$^1$Department of Statistics, University of California, Davis, *hrli@ucdavis.edu; **aaue@ucdavis.edu; †debpaul@ucdavis.edu; ‡jiepeng@ucdavis.edu

$^2$Icahn Institute, Department of Genetics and Genomic Sciences, School of Medicine at Mount Sinai, pei.wang@mssm.edu

We propose a two-sample test for detecting the difference between mean vectors in a high-dimensional regime based on a ridge-regularized Hotelling’s $T^2$. To choose the regularization parameter, a method is derived that aims at maximizing power within a class of local alternatives. We also propose a composite test that combines the optimal tests corresponding to a specific collection of local alternatives. Weak convergence of the stochastic process corresponding to the ridge-regularized Hotelling’s $T^2$ is established and used to derive the cut-off values of the proposed test. Large sample properties are verified for a class of sub-Gaussian distributions. Through an extensive simulation study, the composite test is shown to compare favorably against a host of existing two-sample test procedures in a wide range of settings. The performance of the proposed test procedures is illustrated through an application to a breast cancer data set where the goal is to detect the pathways with different DNA copy number alterations across breast cancer subtypes.

REFERENCES


Key words and phrases. Asymptotic property, covariance matrix, Hotelling’s $T^2$ statistic, hypothesis testing, locally most powerful tests, random matrix theory.


GRID: A VARIABLE SELECTION AND STRUCTURE DISCOVERY METHOD FOR HIGH DIMENSIONAL NONPARAMETRIC REGRESSION

BY FRANCESCO GIORDANO¹,*, SOUMENDRA NATH LAHIRI² AND MARIA LUCIA PARRELLA¹,†

¹Department of Economics and Statistics, University of Salerno, *giordano@unisa.it; †mparrella@unisa.it
²Department of Statistics, North Carolina State University, snlahiri@ncsu.edu

We consider nonparametric regression in high dimensions where only a relatively small subset of a large number of variables are relevant and may have nonlinear effects on the response. We develop methods for variable selection, structure discovery and estimation of the true low-dimensional regression function, allowing any degree of interactions among the relevant variables that need not be specified a-priori. The proposed method, called the GRID, combines empirical likelihood based marginal testing with the local linear estimation machinery in a novel way to select the relevant variables. Further, it provides a simple graphical tool for identifying the low dimensional nonlinear structure of the regression function. Theoretical results establish consistency of variable selection and structure discovery, and also Oracle risk property of the GRID estimator of the regression function, allowing the dimension d of the covariates to grow with the sample size n at the rate $d = O(n^a)$ for any $a \in (0, \infty)$ and the number of relevant covariates r to grow at a rate $r = O(n^\gamma)$ for some $\gamma \in (0, 1)$ under some regularity conditions that, in particular, require finiteness of certain absolute moments of the error variables depending on a. Finite sample properties of the GRID are investigated in a moderately large simulation study.

REFERENCES


Key words and phrases. Empirical likelihood, marginal testing, variable selection consistency.


BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE
STATISTICAL
INFERERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE