Articles

Nonparametric regression using deep neural networks with ReLU activation function.................................JOHANNES SCHMIDT-HIEBER 1875

Discussion of: “Nonparametric regression using deep neural networks with ReLU activation function” ......................... BEHROOZ GHORBANI, SONG MEI, THEODOR MISIAKIEWICZ AND ANDREA MONTANARI 1898

Discussion of: “Nonparametric regression using deep neural networks with ReLU activation function”.................................GITTA KUTYNIOK 1902

Discussion of: “Nonparametric regression using deep neural networks with ReLU activation function”.................................MICHAIL KOHLER AND SOPHIE LANGER 1906

Discussion of: “Nonparametric regression using deep neural networks with ReLU activation function”.................................OHAD SHAMIR 1911


Nonclassical Berry–Esseen inequalities and accuracy of the bootstrap ..........................................................MAYYA ZHILOVA 1922

On the validity of the formal Edgeworth expansion for posterior densities ..................................................JOHN E. KOLASSA AND TODD A. KUFFNER 1940

Model selection for high-dimensional linear regression with dependent observations ..............................................CHING-KANG ING 1959

Optimal estimation of Gaussian mixtures via denoised method of moments ..................................................YIHONG WU AND PENGKUN YANG 1981

Sharp instruments for classifying compliers and generalizing causal effects ..................................................EDWARD H. KENNEDY, SIVARAMAN BALAKRISHNAN AND MAX G’SELL 2008

Empirical risk minimization and complexity of dynamical models ...........................................................................KEVIN MCGOFF AND ANDREW B. NOBEL 2031

Adaptive estimation in structured factor models with applications to overlapping clustering .........................................XIN BING, FLORENTINA BUNEA, YANG NING AND MARTEN WEGKAMP 2055

Partial identifiability of restricted latent class models..................................................................................YUQI GU AND GONGJUN XU 2082

Posterior concentration for Bayesian regression trees and forests ...........................................................................VERONIKA ROČKOVÁ AND STÉPHANIE VAN DER PAS 2108

Double-slicing assisted sufficient dimension reduction for high-dimensional censored data ........................................SHANSHAN DING, WEI QIAN AND LAN WANG 2132

Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors ..................................................................JUDITH ROUSSEAU AND BOTOND SZABO 2155

Convergence rates of variational posterior distributions ..................................................................................FENGSHUO ZHANG AND CHAO GAO 2180

Two-sample hypothesis testing for inhomogeneous random graphs ..............................................................................DEBARGHYA GHOSHDAŠTIDAR, MAURILIO GUTZEIT, ALEXANDRA CARPENTIER AND ULRIKE VON LUXBURG 2208

Beyond HC: More sensitive tests for rare/weak alternatives ..................................................................................THOMAS PORTER AND MICHAEL STEWART 2230

continued
Minimax optimal rates for Mondrian trees and forests
JAOUAD MOURTADA, STÉPHANE GAIFFAS AND ERWAN SCORNET 2253

Identifiability of nonparametric mixture models and Bayes optimal clustering
BRYON ARAGAM, CHEN DAN, ERIC P. XING AND PRADEEP RAVIKUMAR 2277

A test for separability in covariance operators of random surfaces
PRAMITA BAGCHI AND HOLGER DETTE 2303

A general approach for cure models in survival analysis
VALENTIN PATILEA AND INGRID VAN KEILEGOM 2323

Adaptive distributed methods under communication constraints
BOTOND SZABÓ AND HARRY VAN ZANTEN 2347

Bayesian analysis of the covariance matrix of a multivariate normal distribution with a
new class of priors .......................... JAMES O. BERGER, DONGCHU SUN
AND CHENGYUAN SONG 2381

Extending the validity of frequency domain bootstrap methods to general stationary
processes ..................................... MARCO MEYER, EfSTATHIOS PAPARODITIS
AND JENS-PETER KREISS 2404

Minimax estimation of large precision matrices with bandable Cholesky factor
YU LIU AND ZHAO REN 2428

Estimation and inference for precision matrices of nonstationary time series
XIUCAI DING AND ZHOU ZHOU 2455

Isotropic covariance functions on graphs and their edges
ETHAN ANDERES, JESPER MØLLER AND JAKOB G. RASMUSSEN 2478
INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

President: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

Past President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS EDITORS

The Annals of Statistics. Editors: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

The Annals of Applied Statistics. Editor-in-Chief: Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

The Annals of Probability. Editor: Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA


Statistical Science. Editor: Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

The IMS Bulletin. Editor: Vlada Limic, UMR 7501 de l’Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Statistics [ISSN 0090-5364 (print); ISSN 2168-8966 (online)], Volume 48, Number 4, August 2020. Published bimonthly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to The Annals of Statistics, Institute of Mathematical Statistics, Dues and Subscriptions Office, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

Copyright © 2020 by the Institute of Mathematical Statistics
Printed in the United States of America
Consider the multivariate nonparametric regression model. It is shown that estimators based on sparsely connected deep neural networks with ReLU activation function and properly chosen network architecture achieve the minimax rates of convergence (up to \( \log n \)-factors) under a general composition assumption on the regression function. The framework includes many well-studied structural constraints such as (generalized) additive models. While there is a lot of flexibility in the network architecture, the tuning parameter is the sparsity of the network. Specifically, we consider large networks with number of potential network parameters exceeding the sample size. The analysis gives some insights into why multilayer feedforward neural networks perform well in practice. Interestingly, for ReLU activation function the depth (number of layers) of the neural network architectures plays an important role, and our theory suggests that for nonparametric regression, scaling the network depth with the sample size is natural. It is also shown that under the composition assumption wavelet estimators can only achieve suboptimal rates.

REFERENCES


MSC2020 subject classifications. 62G08.
Key words and phrases. Nonparametric regression, multilayer neural networks, ReLU activation function, minimax estimation risk, additive models, wavelets.


DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY BEHROOZ GHORBANI¹, SONG MEI², THEODOR MISIAKIEWICZ⁴ AND ANDREA MONTANARI⁴

¹Department of Electrical Engineering, Stanford University, ghorbani@stanford.edu
²Institute for Computational and Mathematical Engineering, Stanford University, songmei@stanford.edu
³Department of Statistics, Stanford University, misiakie@stanford.edu
⁴Department of Electrical Engineering and Department of Statistics, Stanford University, montanari@stanford.edu

REFERENCES

DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY GITTA KUTYNIOK

Institute of Mathematics, Technische Universität Berlin, kutyniok@math.tu-berlin.de

I would like to congratulate Johannes Schmidt–Hieber on a very interesting paper in which he considers regression functions belonging to the class of so-called compositional functions and analyzes the ability of estimators based on the multivariate nonparametric regression model of deep neural networks to achieve minimax rates of convergence.

In my discussion, I will first regard such a type of result from the general viewpoint of the theoretical foundations of deep neural networks. This will be followed by a discussion from the viewpoint of expressivity, optimization and generalization. Finally, I will consider some specific aspects of the main result.

REFERENCES


MSC2020 subject classifications. 62G08.

Key words and phrases. Deep neural networks, generalization, nonparametric regression.
DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY MICHAEL KOHLER * AND SOPHIE LANGER †

Fachbereich Mathematik, TU Darmstadt, * kohler@mathematik.tu-darmstadt.de; † langer@mathematik.tu-darmstadt.de

REFERENCES


DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY OHAD SHAMIR

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, ohad.shamir@weizmann.ac.il

REFERENCES


REJOINDER: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY JOHANNES SCHMIDT-HIEBER

Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl

REFERENCES


NONCLASSICAL BERRY–ESSEEN INEQUALITIES AND ACCURACY OF THE BOOTSTRAP

BY MAYYA ZHILOVA

School of Mathematics, Georgia Institute of Technology, mzhilova@math.gatech.edu

We study accuracy of bootstrap procedures for estimation of quantiles of a smooth function of a sum of independent sub-Gaussian random vectors. We establish higher-order approximation bounds with error terms depending on a sample size and a dimension explicitly. These results lead to improvements of accuracy of a weighted bootstrap procedure for general log-likelihood ratio statistics. The key element of our proofs of the bootstrap accuracy is a multivariate higher-order Berry–Esseen inequality. We consider a problem of approximation of distributions of two sums of zero mean independent random vectors, such that summands with the same indices have equal moments up to at least the second order. The derived approximation bound is uniform on the sets of all Euclidean balls. The presented approach extends classical Berry–Esseen type inequalities to higher-order approximation bounds. The theoretical results are illustrated with numerical experiments.

REFERENCES


Key words and phrases. Multivariate Berry–Esseen inequality, dependence on dimension, Efron’s bootstrap, weighted bootstrap, multiplier bootstrap, higher-order inference, smooth function model, likelihood-based confidence sets.

ON THE VALIDITY OF THE FORMAL EDELWORTH EXPANSION FOR POSTERIOR DENSITIES

BY JOHN E. KOLASSA¹ AND TODD A. KUFFNER²

¹Department of Statistics and Biostatistics, Rutgers University, kolassa@stat.rutgers.edu
²Department of Mathematics and Statistics, Washington University in St. Louis, kuffner@wustl.edu

We consider a fundamental open problem in parametric Bayesian theory, namely the validity of the formal Edgeworth expansion of the posterior density. While the study of valid asymptotic expansions for posterior distributions constitutes a rich literature, the validity of the formal Edgeworth expansion has not been rigorously established. Several authors have claimed connections of various posterior expansions with the classical Edgeworth expansion, or have simply assumed its validity. Our main result settles this open problem. We also prove a lemma concerning the order of posterior cumulants which is of independent interest in Bayesian parametric theory. The most relevant literature is synthesized and compared to the newly-derived Edgeworth expansions. Numerical investigations illustrate that our expansion has the behavior expected of an Edgeworth expansion, and that it has better performance than the other existing expansion which was previously claimed to be of Edgeworth type.

REFERENCES


Key words and phrases. Edgeworth expansion, higher-order asymptotics, posterior, cumulant expansion.


MODEL SELECTION FOR HIGH-DIMENSIONAL LINEAR REGRESSION WITH DEPENDENT OBSERVATIONS

BY CHING-KANG ING

Institute of Statistics, National Tsing Hua University, cking@stat.nthu.edu.tw

We investigate the prediction capability of the orthogonal greedy algorithm (OGA) in high-dimensional regression models with dependent observations. The rates of convergence of the prediction error of OGA are obtained under a variety of sparsity conditions. To prevent OGA from overfitting, we introduce a high-dimensional Akaike’s information criterion (HDAIC) to determine the number of OGA iterations. A key contribution of this work is to show that OGA, used in conjunction with HDAIC, can achieve the optimal convergence rate without knowledge of how sparse the underlying high-dimensional model is.

REFERENCES


**MSC2020 subject classifications.** Primary 63M30; secondary 62F07, 62F12.

**Key words and phrases.** Best \( m \)-term approximations, high-dimensional Akaike’s information criterion, orthogonal greedy algorithm, sparsity conditions, time series.


OPTIMAL ESTIMATION OF GAUSSIAN MIXTURES VIA DENOISED METHOD OF MOMENTS

BY YIHONG WU¹ AND PENGKUN YANG²

¹Department of Statistics and Data Science, Yale University, yihong.wu@yale.edu
²Department of Electrical Engineering, Princeton University, pengkun@princeton.edu

The method of moments (Philos. Trans. R. Soc. Lond. Ser. A 185 (1894) 71–110) is one of the most widely used methods in statistics for parameter estimation, by means of solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or nonuniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analyses of the method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

This paper considers estimating a $k$-component Gaussian location mixture with a common (possibly unknown) variance parameter. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space (via semidefinite programming) before solving the method of moments equations. Not only does this regularization ensure existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in the Wasserstein distance via polynomial interpolation and majorization techniques, we establish the statistical guarantees and adaptive optimality of the proposed procedure, as well as oracle inequality in misspecified models. These results can also be viewed as provable algorithms for generalized method of moments (Econometrica 50 (1982) 1029–1054) which involves nonconvex optimization and lacks theoretical guarantees.

REFERENCES


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

Key words and phrases. Gaussian mixture, finite mixture model, deconvolution, method of moments, Wasserstein distance, semidefinite programming, moment space, Gauss quadrature, minimax optimality.
SHARP INSTRUMENTS FOR CLASSIFYING COMPLIERS AND GENERALIZING CAUSAL EFFECTS

BY EDWARD H. KENNEDY*, SIVARAMAN BALAKRISHNAN** AND MAX G’SELL†

Department of Statistics and Data Science, Carnegie Mellon University, * edward@stat.cmu.edu; ** siva@stat.cmu.edu; † mgsell@stat.cmu.edu

It is well known that, without restricting treatment effect heterogeneity, instrumental variable (IV) methods only identify “local” effects among compliers, that is, those subjects who take treatment only when encouraged by the IV. Local effects are controversial since they seem to only apply to an unidentified subgroup; this has led many to denounce these effects as having little policy relevance. However, we show that such pessimism is not always warranted: it can be possible to accurately predict who compliers are, and obtain tight bounds on more generalizable effects in identifiable subgroups. We propose methods for doing so and study estimation error and asymptotic properties, showing that these tasks can sometimes be accomplished even with very weak IVs. We go on to introduce a new measure of IV quality called “sharpness,” which reflects the variation in compliance explained by covariates, and captures how well one can identify compliers and obtain tight bounds on identifiable subgroup effects. We develop an estimator of sharpness and show that it is asymptotically efficient under weak conditions. Finally, we explore finite-sample properties via simulation, and apply the methods to study canvassing effects on voter turnout. We propose that sharpness should be presented alongside strength to assess IV quality.

REFERENCES


MSC2020 subject classifications. 62G05, 62H30.
Key words and phrases. Causal inference, instrumental variable, noncompliance, observational study, generalizability.


EMPIRICAL RISK MINIMIZATION AND COMPLEXITY OF DYNAMICAL MODELS

BY KEVIN McGOFF\(^1\) AND ANDREW B. NOBEL\(^2\)

\(^1\)Department of Mathematics and Statistics, University of North Carolina at Charlotte, kmcgoff1@uncc.edu
\(^2\)Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, nobel@email.unc.edu

A dynamical model consists of a continuous self-map \(T : \mathcal{X} \rightarrow \mathcal{X}\) of a compact state space \(\mathcal{X}\) and a continuous observation function \(f : \mathcal{X} \rightarrow \mathbb{R}\). This paper considers the fitting of a parametrized family of dynamical models to an observed real-valued stochastic process using empirical risk minimization. The limiting behavior of the minimum risk parameters is studied in a general setting. We establish a general convergence theorem for minimum risk estimators and ergodic observations. We then study conditions under which empirical risk minimization can effectively separate signal from noise in an additive observational noise model. The key condition in the latter results is that the family of dynamical models has limited complexity, which is quantified through a notion of entropy for families of infinite sequences that connects covering number based entropies with topological entropy studied in dynamical systems. We establish close connections between entropy and limiting average mean widths for stationary processes, and discuss several examples of dynamical models.

REFERENCES


MSC2020 subject classifications. Primary 62M09.
Key words and phrases. Empirical risk minimization, dynamical models, joinings, topological entropy.


ADAPTIVE ESTIMATION IN STRUCTURED FACTOR MODELS WITH APPLICATIONS TO OVERLAPPING CLUSTERING

BY XIN BING1,*, FLORENTINA BUNEA1,**, YANG NING1,† AND MARTEN WEGKAMP2

1Department of Statistics and Data Science, Cornell University, xb43@cornell.edu; fb238@cornell.edu; yn265@cornell.edu
2Department of Mathematics & Department of Statistics and Data Science, Cornell University, mhw73@cornell.edu

This work introduces a novel estimation method, called LOVE, of the entries and structure of a loading matrix $A$ in a latent factor model $X = AZ + E$, for an observable random vector $X \in \mathbb{R}^p$, with correlated unobservable factors $Z \in \mathbb{R}^K$, with $K$ unknown, and uncorrelated noise $E$. Each row of $A$ is scaled, and allowed to be sparse. In order to identify the loading matrix $A$, we require the existence of pure variables, which are components of $X$ that are associated, via $A$, with one and only one latent factor. Despite the fact that the number of factors $K$, the number of the pure variables and their location are all unknown, we only require a mild condition on the covariance matrix of $Z$, and a minimum of only two pure variables per latent factor to show that $A$ is uniquely defined, up to signed permutations. Our proofs for model identifiability are constructive, and lead to our novel estimation method of the number of factors and of the set of pure variables, from a sample of size $n$ of observations on $X$. This is the first step of our LOVE algorithm, which is optimization-free, and has low computational complexity of order $p^2$. The second step of LOVE is an easily implementable linear program that estimates $A$. We prove that the resulting estimator is near minimax rate optimal for $A$, with respect to the $\| \cdot \|_{\infty, q}$ loss, for $q \geq 1$, up to logarithmic factors in $p$, and that it can be minimax-rate optimal in many cases of interest.

The model structure is motivated by the problem of overlapping variable clustering, ubiquitous in data science. We define the population level clusters as groups of those components of $X$ that are associated, via the matrix $A$, with the same unobservable latent factor, and multifactor association is allowed. Clusters are respectively anchored by the pure variables, and form overlapping subgroups of the $p$-dimensional random vector $X$. The Latent model approach to OVERlapping clustering is reflected in the name of our algorithm, LOVE.

The third step of LOVE estimates the clusters from the support of the columns of the estimated $A$. We guarantee cluster recovery with zero false positive proportion, and with false negative proportion control. The practical relevance of LOVE is illustrated through the analysis of a RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.

REFERENCES


Key words and phrases. Overlapping clustering, latent model, identification, high-dimensional estimation, minimax estimation, pure variables, group recovery, support recovery, sparse loading matrix, matrix factorization, adaptive estimation.
PARTIAL IDENTIFIABILITY OF RESTRICTED LATENT CLASS MODELS

BY YUQI GU* AND GONGJUN XU**

Department of Statistics, University of Michigan, *yuqigu@umich.edu; **gongjun@umich.edu

Latent class models have wide applications in social and biological sciences. In many applications, prespecified restrictions are imposed on the parameter space of latent class models, through a design matrix, to reflect practitioners’ assumptions about how the observed responses depend on subjects’ latent traits. Though widely used in various fields, such restricted latent class models suffer from nonidentifiability due to their discreteness nature and complex structure of restrictions. This work addresses the fundamental identifiability issue of restricted latent class models by developing a general framework for strict and partial identifiability of the model parameters. Under correct model specification, the developed identifiability conditions only depend on the design matrix and are easily checkable, which provide useful practical guidelines for designing statistically valid diagnostic tests. Furthermore, the new theoretical framework is applied to establish, for the first time, identifiability of several designs from cognitive diagnosis applications.

REFERENCES


MSC2020 subject classifications. 62E10, 62P15.

Key words and phrases. Identifiability, restricted latent class models, cognitive diagnosis, $Q$-matrix.


POSTERIOR CONCENTRATION FOR BAYESIAN REGRESSION TREES AND FORESTS

BY VERONIKA ROČKOVÁ¹ AND STÉPHANIE VAN DER PAS²

¹Booth School of Business, University of Chicago, veronika.rockova@chicagobooth.edu
²Mathematics Institute, Leiden University, svdpas@math.leidenuniv.nl

Since their inception in the 1980s, regression trees have been one of the more widely used nonparametric prediction methods. Tree-structured methods yield a histogram reconstruction of the regression surface, where the bins correspond to terminal nodes of recursive partitioning. Trees are powerful, yet susceptible to overfitting. Strategies against overfitting have traditionally relied on pruning greedily grown trees. The Bayesian framework offers an alternative remedy against overfitting through priors. Roughly speaking, a good prior charges smaller trees where overfitting does not occur. While the consistency of random histograms, trees and their ensembles has been studied quite extensively, the theoretical understanding of the Bayesian counterparts has been missing. In this paper, we take a step toward understanding why/when do Bayesian trees and forests not overfit. To address this question, we study the speed at which the posterior concentrates around the true smooth regression function. We propose a spike-and-tree variant of the popular Bayesian CART prior and establish new theoretical results showing that regression trees (and forests) (a) are capable of recovering smooth regression surfaces (with smoothness not exceeding one), achieving optimal rates up to a log factor, (b) can adapt to the unknown level of smoothness and (c) can perform effective dimension reduction when p > n. These results provide a piece of missing theoretical evidence explaining why Bayesian trees (and additive variants thereof) have worked so well in practice.

REFERENCES


MSC2020 subject classifications. 62G08, 62G20.

Key words and phrases. Additive regression, asymptotic minimaxity, BART, Bayesian CART, posterior concentration, recursive partitioning, regression trees.
DOUBLE-SLICING ASSISTED SUFFICIENT DIMENSION REDUCTION FOR HIGH-DIMENSIONAL CENSORED DATA

BY SHANSHAN DING1,*, WEI QIAN1,** AND LAN WANG2

1Department of Applied Economics and Statistics, University of Delaware, *sding@udel.edu; **weiqian@udel.edu
2Department of Management Science, University of Miami, lanwang@mbs.miami.edu

This paper provides a unified framework and an efficient algorithm for analyzing high-dimensional survival data under weak modeling assumptions. In particular, it imposes neither parametric distributional assumption nor linear regression assumption. It only assumes that the survival time \( T \) depends on a high-dimensional covariate vector \( X \) through low-dimensional linear combinations of covariates \( \Gamma^T X \). The censoring time is allowed to be conditionally independent of the survival time given the covariates. This general framework includes many popular parametric and semiparametric survival regression models as special cases. The proposed algorithm produces a number of practically useful outputs with theoretical guarantees, including a consistent estimate of the sufficient dimension reduction subspace of \( T \mid X \), a uniformly consistent Kaplan–Meier-type estimator of the conditional distribution function of \( T \) and a consistent estimator of the conditional quantile survival time. Our asymptotic results significantly extend the classical theory of sufficient dimension reduction for censored data (particularly that of Li, Wang and Chen in Ann. Statist. 27 (1999) 1–23) and the celebrated nonparametric Kaplan–Meier estimator to the setting where the number of covariates \( p \) diverges exponentially fast with the sample size \( n \). We demonstrate the promising performance of the proposed new estimators through simulations and a real data example.

REFERENCES


Key words and phrases. Central subspace, sufficient dimension reduction, variable selection, censored data, ultrahigh dimension, nonparametric estimation.
https://doi.org/10.1198/016214506000000735
ASYMPTOTIC FREQUENTIST COVERAGE PROPERTIES OF BAYESIAN CREDIBLE SETS FOR SIEVE PRIORS

BY JUDITH ROUSSEAU$^1$ AND BOTOND SZABO$^2$

$^1$Statistics Department, Oxford University, rousseau@ceremade.dauphine.fr
$^2$Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl

We investigate the frequentist coverage properties of (certain) Bayesian credible sets in a general, adaptive, nonparametric framework. It is well known that the construction of adaptive and honest confidence sets is not possible in general. To overcome this problem (in context of sieve type of priors), we introduce an extra assumption on the functional parameters, the so-called “general polished tail” condition. We then show that under standard assumptions, both the hierarchical and empirical Bayes methods, result in honest confidence sets for sieve type of priors in general settings and we characterize their size. We apply the derived abstract results to various examples, including the nonparametric regression model, density estimation using exponential families of priors, density estimation using histogram priors and the nonparametric classification model, for which we show that their size is near minimax adaptive with respect to the considered specific pseudometrics.

REFERENCES


MSC2020 subject classifications. Primary 62G20, 62G05; secondary 62G08, 62G07.

Key words and phrases. Uncertainty quantification, coverage, posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, classification, sieve prior.
CONVERGENCE RATES OF VARIATIONAL POSTERIOR DISTRIBUTIONS

BY FENGSHUO ZHANG* AND CHAO GAO†

Department of Statistics, University of Chicago, *fengshuo@galton.uchicago.edu; †chaogao@galton.uchicago.edu

We study convergence rates of variational posterior distributions for non-parametric and high-dimensional inference. We formulate general conditions on prior, likelihood and variational class that characterize the convergence rates. Under similar “prior mass and testing” conditions considered in the literature, the rate is found to be the sum of two terms. The first term stands for the convergence rate of the true posterior distribution, and the second term is contributed by the variational approximation error. For a class of priors that admit the structure of a mixture of product measures, we propose a novel prior mass condition, under which the variational approximation error of the mean-field class is dominated by convergence rate of the true posterior. We demonstrate the applicability of our general results for various models, prior distributions and variational classes by deriving convergence rates of the corresponding variational posteriors.

REFERENCES


MSC2020 subject classifications. Primary 62C10; secondary 62F15.

Key words and phrases. Posterior contraction, mean-field variational inference, density estimation, Gaussian sequence model, piecewise constant model, empirical Bayes.
TWO-SAMPLE HYPOTHESIS TESTING FOR INHOMOGENEOUS RANDOM GRAPHS

BY DEBARGHYA GHOSHDASTIDAR1,*, MAURILIO GUTZEIT2,†, ALEXANDRA CARPENTIER2,‡ AND ULRIKE VON LUXBURG1,**

1Department of Computer Science, University of Tübingen, * debarghya.ghoshdastidar@uni-tuebingen.de;
**luxburg@informatik.uni-tuebingen.de
2Faculty of Mathematics, Otto von Guericke University Magdeburg, † maurilio.gutzeit@ovgu.de;
‡alexandra.carpentier@ovgu.de

The study of networks leads to a wide range of high-dimensional inference problems. In many practical applications, one needs to draw inference from one or few large sparse networks. The present paper studies hypothesis testing of graphs in this high-dimensional regime, where the goal is to test between two populations of inhomogeneous random graphs defined on the same set of \( n \) vertices. The size of each population \( m \) is much smaller than \( n \), and can even be a constant as small as 1. The critical question in this context is whether the problem is solvable for small \( m \).

We answer this question from a minimax testing perspective. Let \( P, Q \) be the population adjacencies of two sparse inhomogeneous random graph models, and \( d \) be a suitably defined distance function. Given a population of \( m \) graphs from each model, we derive minimax separation rates for the problem of testing \( P = Q \) against \( d(P, Q) > \rho \). We observe that if \( m \) is small, then the minimax separation is too large for some popular choices of \( d \), including total variation distance between corresponding distributions. This implies that some models that are widely separated in \( d \) cannot be distinguished for small \( m \), and hence, the testing problem is generally not solvable in these cases.

We also show that if \( m > 1 \), then the minimax separation is relatively small if \( d \) is the Frobenius norm or operator norm distance between \( P \) and \( Q \). For \( m = 1 \), only the latter distance provides small minimax separation. Thus, for these distances, the problem is solvable for small \( m \). We also present near-optimal two-sample tests in both cases, where tests are adaptive with respect to sparsity level of the graphs.

REFERENCES


MSC2020 subject classifications. Primary 62H15; secondary 62C20, 05C80, 60B20.
Key words and phrases. Two-sample test, inhomogeneous Erdős–Rényi model, minimax testing.


BEYOND HC: MORE SENSITIVE TESTS FOR RARE/WEAK ALTERNATIVES

BY THOMAS PORTER\(^1\) AND MICHAEL STEWART\(^2\)

In memory of Peter Hall

\(^1\)School of Mathematics and Statistics, University of Melbourne, thomas.porter@outlook.com.au
\(^2\)School of Mathematics and Statistics F07, University of Sydney, michael.stewart@sydney.edu.au

Higher criticism (HC) is a popular method for large-scale inference problems based on identifying unusually high proportions of small \(p\)-values. It has been shown to enjoy a lower-order optimality property in a simple normal location mixture model which is shared by the ‘tailor-made’ parametric generalised likelihood ratio test (GLRT) for the same model; however, HC has also been shown to perform well outside this ‘narrow’ model.

We develop a higher-order framework for analysing the power of these and similar procedures, which reveals the perhaps unsurprising fact that the GLRT enjoys an edge in power over HC for the normal location mixture model. We also identify a similar parametric mixture model to which HC is similarly ‘tailor-made’ and show that the situation is (at least partly) reversed there. We also show that in the normal location mixture model a procedure based on the empirical moment-generating function enjoys the same local power properties as the GLRT and may be recommended as an easy to implement (and interpret), complementary procedure to HC. Some other practical advice regarding the implementation of these procedures is provided. Finally, we provide some simulation results to help interpret our theoretical findings.

REFERENCES


Key words and phrases. Higher criticism, sparse normal means, phi-divergence, multiple comparisons, mixture model.
MINIMAX OPTIMAL RATES FOR MONDRIAN TREES AND FORESTS

BY JAOUAD MOURTADA 1,*, STÉPHANE GAÏFFAS 2 AND ERWAN SCORNET 1,**

1CMAP, École polytechnique, *jaouad.mourtada@polytechnique.edu; **erwan.scornet@polytechnique.edu
2LPMA—Université Paris Diderot, stephane.gaiffas@lpsm.paris

Introduced by Breiman (Mach. Learn. 45 (2001) 5–32), Random Forests are widely used classification and regression algorithms. While being initially designed as batch algorithms, several variants have been proposed to handle online learning. One particular instance of such forests is the Mondrian forest (In Adv. Neural Inf. Process. Syst. (2014) 3140–3148; In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) (2016)), whose trees are built using the so-called Mondrian process, therefore allowing to easily update their construction in a streaming fashion. In this paper we provide a thorough theoretical study of Mondrian forests in a batch learning setting, based on new results about Mondrian partitions. Our results include consistency and convergence rates for Mondrian trees and forests, that turn out to be minimax optimal on the set of \( s \)-Hölder function with \( s \in (0, 1] \) (for trees and forests) and \( s \in (1, 2] \) (for forests only), assuming a proper tuning of their complexity parameter in both cases. Furthermore, we prove that an adaptive procedure (to the unknown \( s \in (0, 2] \)) can be constructed by combining Mondrian forests with a standard model aggregation algorithm. These results are the first demonstrating that some particular random forests achieve minimax rates in arbitrary dimension. Owing to their remarkably simple distributional properties, which lead to minimax rates, Mondrian trees are a promising basis for more sophisticated yet theoretically sound random forests variants.

REFERENCES


IDENTIFIABILITY OF NONPARAMETRIC MIXTURE MODELS AND BAYES OPTIMAL CLUSTERING

BY BRYON ARAGAM¹, CHEN DAN²,*, ERIC P. XING²,** AND PRADEEP RAVIKUMAR²,†

¹Booth School of Business, University of Chicago, bryon@chicagobooth.edu
²Machine Learning Department, School of Computer Science, Carnegie Mellon University, *cdan@andrew.cmu.edu; **epxing@cs.cmu.edu; †pradeepr@cs.cmu.edu

Motivated by problems in data clustering, we establish general conditions under which families of nonparametric mixture models are identifiable by introducing a novel framework involving clustering overfitted parametric (i.e., misspecified) mixture models. These identifiability conditions generalize existing conditions in the literature and are flexible enough to include, for example, mixtures of infinite Gaussian mixtures. In contrast to the recent literature, we allow for general nonparametric mixture components and instead impose regularity assumptions on the underlying mixing measure. As our primary application we apply these results to partition-based clustering, generalizing the notion of a Bayes optimal partition from classical parametric model-based clustering to nonparametric settings. Furthermore, this framework is constructive, so that it yields a practical algorithm for learning identified mixtures, which is illustrated through several examples on real data. The key conceptual device in the analysis is the convex, metric geometry of probability measures on metric spaces and its connection to the Wasserstein convergence of mixing measures. The result is a flexible framework for nonparametric clustering with formal consistency guarantees.

REFERENCES


MSC2020 subject classifications. Primary 62G05, 62H30; secondary 62H12.

Key words and phrases. Mixture models, nonparametric statistics, identifiability, clustering, Bayes optimal partition.


A TEST FOR SEPARABILITY IN COVARIANCE OPERATORS OF RANDOM SURFACES

BY PRAMITA BAGCHI¹ AND HOLGER DETTE²

¹Department of Statistics, George Mason University
²Fakultät für Mathematik, Ruhr-Universität Bochum, holger.dette@rub.de

The assumption of separability is a simplifying and very popular assumption in the analysis of spatiotemporal or hypersurface data structures. It is often made in situations where the covariance structure cannot be easily estimated, for example, because of a small sample size or because of computational storage problems. In this paper we propose a new and very simple test to validate this assumption. Our approach is based on a measure of separability which is zero in the case of separability and positive otherwise. We derive the asymptotic distribution of a corresponding estimate under the null hypothesis and the alternative and develop an asymptotic and a bootstrap test which are very easy to implement. In particular, our approach does neither require projections on subspaces generated by the eigenfunctions of the covariance operator nor distributional assumptions as recently used by (Ann. Statist. 45 (2017) 1431–1461) and (Biometrika 104 425–437) to construct tests for separability. We investigate the finite sample performance by means of a simulation study and also provide a comparison with the currently available methodology. Finally, the new procedure is illustrated analyzing a data example.

REFERENCES


Key words and phrases. Functional data, minimum distance, separability, space-time processes, surface data structures.


A GENERAL APPROACH FOR CURE MODELS IN SURVIVAL ANALYSIS

BY VALENTIN PATILEA 1 AND INGRID VAN KEILEGOM 2

1 CREST (UMR 9194), Ensai, Univ Rennes, patilea@ensai.fr
2 ORSTAT, KU Leuven, ingrid.vankeilegom@kuleuven.be

In survival analysis it often happens that some subjects under study do not experience the event of interest; they are considered to be “cured.” The population is thus a mixture of two subpopulations, one of cured subjects and one of “susceptible” subjects. We propose a novel approach to estimate a mixture cure model when covariates are present and the lifetime is subject to random right censoring. We work with a parametric model for the cure proportion, while the conditional survival function of the uncured subjects is unspecified. The approach is based on an inversion which allows us to write the survival function as a function of the distribution of the observable variables. This leads to a very general class of models which allows a flexible and rich modeling of the conditional survival function. We show the identifiability of the proposed model as well as the consistency and the asymptotic normality of the model parameters. We also consider in more detail the case where kernel estimators are used for the nonparametric part of the model. The new estimators are compared with the estimators from a Cox mixture cure model via simulations. Finally, we apply the new model on a medical data set.

REFERENCES


MSC2020 subject classifications. Primary 62N01, 62N02; secondary 62F12, 62G05.

Key words and phrases. Asymptotic normality, bootstrap, kernel smoothing, logistic regression, mixture cure model, semiparametric model.


ADAPTIVE DISTRIBUTED METHODS UNDER COMMUNICATION CONSTRAINTS

BY BONTOND SZABÓ1 AND HARRY VAN ZANTEN2

1Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl
2Department of Mathematics, Faculty of Sciences, VU Amsterdam, j.h.van.zanten@vu.nl

We study estimation methods under communication constraints in a distributed version of the nonparametric random design regression model. We derive minimax lower bounds and exhibit methods that attain those bounds. Moreover, we show that adaptive estimation is possible in this setting.

REFERENCES


Key words and phrases. Distributed computation, minimax rates, adaptation, nonparametric regression, communication constraints.


BAYESIAN ANALYSIS OF THE COVARIANCE MATRIX OF A MULTIVARIATE NORMAL DISTRIBUTION WITH A NEW CLASS OF PRIORS

BY JAMES O. BERGER¹, DONGCHU SUN² AND CHENGYUAN SONG*³

¹Department of Statistical Science, Duke University, berger@stat.duke.edu
²Department of Statistics, University of Nebraska-Lincoln, sund9@unl.edu
³School of Statistics, East China Normal University, songchengyuanchina@163.com

Bayesian analysis for the covariance matrix of a multivariate normal distribution has received a lot of attention in the last two decades. In this paper, we propose a new class of priors for the covariance matrix, including both inverse Wishart and reference priors as special cases. The main motivation for the new class is to have available priors—both subjective and objective—that do not “force eigenvalues apart,” which is a criticism of inverse Wishart and Jeffreys priors. Extensive comparison of these “shrinkage priors” with inverse Wishart and Jeffreys priors is undertaken, with the new priors seeming to have considerably better performance. A number of curious facts about the new priors are also observed, such as that the posterior distribution will be proper with just three vector observations from the multivariate normal distribution—regardless of the dimension of the covariance matrix—and that useful inference about features of the covariance matrix can be possible. Finally, a new MCMC algorithm is developed for this class of priors and is shown to be computationally effective for matrices of up to 100 dimensions.

REFERENCES

EXTENDING THE VALIDITY OF FREQUENCY DOMAIN BOOTSTRAP METHODS TO GENERAL STATIONARY PROCESSES

BY MARCO MEYER1,⁎, EFSTATHIOS PAPARODITIS2 AND JENS-PETER KREISS1,**

1Inst. f. Math. Stochastik, Technische Universität Braunschweig, *marco.meyer@tu-bs.de; **j.kreiss@tu-bs.de
2Department of Mathematics and Statistics, University of Cyprus, stathisp@ucy.ac.cy

Existing frequency domain methods for bootstrapping time series have a limited range. Essentially, these procedures cover the case of linear time series with independent innovations, and some even require the time series to be Gaussian. In this paper we propose a new frequency domain bootstrap method—the hybrid periodogram bootstrap (HPB)—which is consistent for a much wider range of stationary, even nonlinear, processes and which can be applied to a large class of periodogram-based statistics. The HPB is designed to combine desirable features of different frequency domain techniques while overcoming their respective limitations. It is capable to imitate the weak dependence structure of the periodogram by invoking the concept of convolved subsampling in a novel way that is tailor-made for periodograms. We show consistency for the HPB procedure for a general class of stationary time series, ranging clearly beyond linear processes, and for spectral means and ratio statistics on which we mainly focus. The finite sample performance of the new bootstrap procedure is illustrated via simulations.

REFERENCES


Key words and phrases. Bootstrap, periodogram, spectral means, stationary processes.


MINIMAX ESTIMATION OF LARGE PRECISION MATRICES WITH BANDABLE CHOLESKY FACTOR

BY YU LIU* AND ZHAO REN**

Department of Statistics, University of Pittsburgh, *yul125@pitt.edu, **zren@pitt.edu

The last decade has witnessed significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this paper we focus on two commonly used types of parameter spaces and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found. Two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed. For the operator norm our optimal procedure is based on a novel local cropping estimator, targeting on all principle submatrices of the precision matrix, while for the Frobenius norm our optimal procedure relies on a delicate regression-based thresholding rule. Lepski’s method is considered to achieve optimal adaptation. We further establish rate optimality in the nonparanormal model. Numerical studies are carried out to confirm our theoretical findings.

REFERENCES


Key words and phrases. Optimal rate of convergence, precision matrix, local cropping, Cholesky factor, minimax lower bound, thresholding, operator norm, Frobenius norm and adaptive estimation.


ESTIMATION AND INFERENCE FOR PRECISION MATRICES OF NONSTATIONARY TIME SERIES

BY XIUCAI DING* AND ZHOU ZHOU**

Department of Statistics, University of Toronto, *xiucai.ding@mail.utoronto.ca; **zhou@utstat.toronto.edu

We consider the estimation of and inference on precision matrices of a rich class of univariate locally stationary linear and nonlinear time series, assuming that only one realization of the time series is observed. Using a Cholesky decomposition technique, we show that the precision matrices can be directly estimated via a series of least squares linear regressions with smoothly time-varying coefficients. The method of sieves is utilized for the estimation and is shown to be optimally adaptive in terms of estimation accuracy and efficient in terms of computational complexity. We establish an asymptotic theory for a class of $L^2$ tests based on the nonparametric sieve estimators. The latter are used for testing whether the precision matrices are diagonal or banded. A Gaussian approximation result is established for a wide class of quadratic forms of nonstationary and possibly nonlinear processes of diverging dimensions which is of interest by itself.

REFERENCES


Key words and phrases. Nonstationary time series, precision matrices, Cholesky decomposition, sieve estimation, high-dimensional Gaussian approximation, random matrices, white noise and bandedness tests.
ISOTROPIC COVARIANCE FUNCTIONS ON GRAPHS AND THEIR EDGES

BY ETHAN ANDERES1, JESPER MØLLER2,* AND JAKOB G. RASMUSSEN2,**

1 Department of Statistics, University of California at Davis, anderes@ucdavis.edu
2 Department of Mathematical Sciences, Aalborg University, *jm@math.aau.dk; **jgr@math.aau.dk

We develop parametric classes of covariance functions on linear networks and their extension to graphs with Euclidean edges, that is, graphs with edges viewed as line segments or more general sets with a coordinate system allowing us to consider points on the graph which are vertices or points on an edge. Our covariance functions are defined on the vertices and edge points of these graphs and are isotropic in the sense that they depend only on the geodesic distance or on a new metric called the resistance metric (which extends the classical resistance metric developed in electrical network theory on the vertices of a graph to the continuum of edge points). We discuss the advantages of using the resistance metric in comparison with the geodesic metric as well as the restrictions these metrics impose on the investigated covariance functions. In particular, many of the commonly used isotropic covariance functions in the spatial statistics literature (the power exponential, Matérn, generalized Cauchy and Dagum classes) are shown to be valid with respect to the resistance metric for any graph with Euclidean edges, whilst they are only valid with respect to the geodesic metric in more special cases.

REFERENCES


Key words and phrases. Geodesic metric, linear network, parametric classes of covariance functions, reproducing kernel Hilbert space, resistance metric, restricted covariance function properties.


BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE
STATISTICAL
INFEERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE

IMS members get a
40% discount
Order your copy now from
 cambridge.org/ims