## Vol. 45, No. 5—October 2017

### Articles

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Local” vs. “global” parameters—breaking the Gaussian complexity barrier</td>
<td>SHAHAR MENDELSON</td>
<td>1835</td>
</tr>
<tr>
<td>Confounder adjustment in multiple hypothesis testing</td>
<td>JINGSHU WANG, QINGYUAN ZHAO, TREvor HASTIE AND ART B. OWEN</td>
<td>1863</td>
</tr>
<tr>
<td>Gaussian approximation for high dimensional time series</td>
<td>DANNA ZHANG AND WEI BIAO WU</td>
<td>1895</td>
</tr>
<tr>
<td>Detection and feature selection in sparse mixture models</td>
<td>NICOLAS VERZELEN AND ERY ARIAS-Castro</td>
<td>1920</td>
</tr>
<tr>
<td>Minimax estimation of a functional on a structured high-dimensional model</td>
<td>JAMES M. ROBINS, LINGLING LI, RAJARSHI MUKHERJEE, ERY TCHETGEN TCHETGEN AND AAD VAN DER VAART</td>
<td>1951</td>
</tr>
<tr>
<td>Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations</td>
<td>MORTEN ÖVERGAARD, ERIK THORLUND PARNER AND JAN PEDERSEN</td>
<td>1988</td>
</tr>
<tr>
<td>Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors</td>
<td>LANCELOT F. JAMES</td>
<td>2016</td>
</tr>
<tr>
<td>Information-regret compromise in covariate-adaptive treatment allocation</td>
<td>ASYA METEKLINA AND LUC PRONZATO</td>
<td>2046</td>
</tr>
<tr>
<td>Sparse CCA: Adaptive estimation and computational barriers</td>
<td>CHAO GAO, ZONGMING MA AND HARRISON H. ZHOU</td>
<td>2074</td>
</tr>
<tr>
<td>Optimal designs for dose response curves with common parameters</td>
<td>CHRYSTEL FELLER, KIRSTEN SCHORNING, HOLGER DETTE, GEORGINA BERMANN AND BJÖRN BORNKAMP</td>
<td>2102</td>
</tr>
<tr>
<td>False discoveries occur early on the Lasso path</td>
<td>WEIJIE SU, MALGORZATA BOGDAN AND EMMANUEL CANDÈS</td>
<td>2133</td>
</tr>
<tr>
<td>Phase transitions for high dimensional clustering and related problems</td>
<td>JIAshun JIN, ZHENG TRACy KE AND WANJIE WANG</td>
<td>2151</td>
</tr>
<tr>
<td>Bayesian detection of image boundaries</td>
<td>MENG LI AND SUBHASHIs GHOSAL</td>
<td>2190</td>
</tr>
<tr>
<td>Spectrum estimation from samples</td>
<td>WEIHAO KONG AND GREGORY VALIANT</td>
<td>2218</td>
</tr>
<tr>
<td>On the contraction properties of some high-dimensional quasi-posterior distributions</td>
<td>YVES A. ATCHADÉ</td>
<td>2248</td>
</tr>
<tr>
<td>Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients</td>
<td>YING ZHU</td>
<td>2274</td>
</tr>
</tbody>
</table>
"LOCAL" VS. "GLOBAL" PARAMETERS—BREAKING THE
GAUSSIAN COMPLEXITY BARRIER

BY SHAHAR MENDELSON1

Technion—Israel Institute of Technology

We show that if $F$ is a convex class of functions that is $L$-sub-Gaussian, the error rate of learning problems generated by independent noise is equivalent to a fixed point determined by "local" covering estimates of the class (i.e., the covering number at a specific level), rather than by the Gaussian average, which takes into account the structure of $F$ at an arbitrarily small scale. To that end, we establish new sharp upper and lower estimates on the error rate in such learning problems.

REFERENCES


MSC2010 subject classifications. 62G08, 62C20, 60G15.

Key words and phrases. Error rates, Gaussian averages, covering numbers.


CONFOUNDER ADJUSTMENT IN MULTIPLE HYPOTHESIS TESTING

BY JINGSHU WANG*, QINGYUAN ZHAO*, TREVOR HASTIE†,1 AND ART B. OWEN†,2

University of Pennsylvania* and Stanford University†

We consider large-scale studies in which thousands of significance tests are performed simultaneously. In some of these studies, the multiple testing procedure can be severely biased by latent confounding factors such as batch effects and unmeasured covariates that correlate with both primary variable(s) of interest (e.g., treatment variable, phenotype) and the outcome. Over the past decade, many statistical methods have been proposed to adjust for the confounders in hypothesis testing. We unify these methods in the same framework, generalize them to include multiple primary variables and multiple nuisance variables, and analyze their statistical properties. In particular, we provide theoretical guarantees for RUV-4 [Gagnon-Bartsch, Jacob and Speed (2013)] and LEAPP [Ann. Appl. Stat. 6 (2012) 1664–1688], which correspond to two different identification conditions in the framework: the first requires a set of “negative controls” that are known a priori to follow the null distribution; the second requires the true nonnulls to be sparse. Two different estimators which are based on RUV-4 and LEAPP are then applied to these two scenarios. We show that if the confounding factors are strong, the resulting estimators can be asymptotically as powerful as the oracle estimator which observes the latent confounding factors. For hypothesis testing, we show the asymptotic $z$-tests based on the estimators can control the type I error. Numerical experiments show that the false discovery rate is also controlled by the Benjamini–Hochberg procedure when the sample size is reasonably large.

REFERENCES


Key words and phrases. Empirical null, surrogate variable analysis, unwanted variation, batch effect, robust regression.


GAUSSIAN APPROXIMATION FOR HIGH DIMENSIONAL TIME SERIES

BY DANNA ZHANG AND WEI BIAO WU

University of California, San Diego and University of Chicago

We consider the problem of approximating sums of high dimensional stationary time series by Gaussian vectors, using the framework of functional dependence measure. The validity of the Gaussian approximation depends on the sample size $n$, the dimension $p$, the moment condition and the dependence of the underlying processes. We also consider an estimator for long-run covariance matrices and study its convergence properties. Our results allow constructing simultaneous confidence intervals for mean vectors of high-dimensional time series with asymptotically correct coverage probabilities. As an application, we propose a Kolmogorov–Smirnov-type statistic for testing distributions of high-dimensional time series.

REFERENCES


MSC2010 subject classifications. Primary 62M10; secondary 62E17.

Key words and phrases. Gaussian approximation, high-dimensional time series, Kolmogorov–Smirnov test, long run covariance matrix, simultaneous inference.


DETECTION AND FEATURE SELECTION IN SPARSE MIXTURE MODELS

BY NICOLAS VERZELEN AND ERY ARIAS-CASTRO

INRA and University of California, San Diego

We consider Gaussian mixture models in high dimensions, focusing on the twin tasks of detection and feature selection. Under sparsity assumptions on the difference in means, we derive minimax rates for the problems of testing and of variable selection. We find these rates to depend crucially on the knowledge of the covariance matrices and on whether the mixture is symmetric or not. We establish the performance of various procedures, including the top sparse eigenvalue of the sample covariance matrix (popular in the context of Sparse PCA), as well as new tests inspired by the normality tests of Malkovich and Afifi [J. Amer. Statist. Assoc. 68 (1973) 176–179].

REFERENCES


MSC2010 subject classifications. 62H30, 62H15.

Key words and phrases. Gaussian mixture models, detection of mixtures, feature selection for mixtures, sparse mixture models, the sparse eigenvalue problem, projection tests based on moments.


MINIMAX ESTIMATION OF A FUNCTIONAL ON A STRUCTURED HIGH-DIMENSIONAL MODEL

BY JAMES M. ROBINS*, LINGLING LI†, RAJARSHI MUKHERJEE‡, ERIC TCHETGEN TCHETGEN* AND AAD VAN DER VAART§

Harvard T. H. Chan School of Public Health,* Sanofi Genzyme,† Stanford University‡ and Universiteit Leiden§

We introduce a new method of estimation of parameters in semiparametric and nonparametric models. The method employs $U$-statistics that are based on higher-order influence functions of the parameter of interest, which extend ordinary linear influence functions, and represent higher derivatives of this parameter. For parameters for which the representation cannot be perfect the method often leads to a bias-variance trade-off, and results in estimators that converge at a slower than $\sqrt{n}$-rate. In a number of examples, the resulting rate can be shown to be optimal. We are particularly interested in estimating parameters in models with a nuisance parameter of high dimension or low regularity, where the parameter of interest cannot be estimated at $\sqrt{n}$-rate, but we also consider efficient $\sqrt{n}$-estimation using novel nonlinear estimators. The general approach is applied in detail to the example of estimating a mean response when the response is not always observed.

REFERENCES


Key words and phrases. Nonlinear functional, nonparametric estimation, $U$-statistic, influence function, tangent space.


ASYMPTOTIC THEORY OF GENERALIZED ESTIMATING EQUATIONS BASED ON JACK-KNIFE PSEUDO-OBSERVATIONS

BY MORTEN OVERGAARD, ERIK THORLUND PARNER AND JAN PEDERSEN

Aarhus University

A general asymptotic theory of estimates from estimating functions based on jack-knife pseudo-observations is established by requiring that the underlying estimator can be expressed as a smooth functional of the empirical distribution. Using results in p-variation norms, the theory is applied to important estimators from time-to-event analysis, namely the Kaplan–Meier estimator and the Aalen–Johansen estimator in a competing risks model, and the corresponding estimators of restricted mean survival and cause-specific lifetime lost. Under an assumption of completely independent censorings, this allows for estimating parameters in regression models of survival, cumulative incidences, restricted mean survival, and cause-specific lifetime lost. Considering estimators as functionals and applying results in p-variation norms is apparently an excellent way of studying the asymptotics of such estimators.

REFERENCES


MSC2010 subject classifications. Primary 62N02; secondary 62F12, 62J12.

Key words and phrases. Pseudo-values, functional differentiability, von Mises expansion, pseudo-observation method, p-variation, U-statistics.


BAYESIAN POISSON CALCULUS FOR LATENT FEATURE MODELING VIA GENERALIZED INDIAN BUFFET PROCESS PRIORS

BY LANCELOT F. JAMES

Hong Kong University of Science and Technology

Statistical latent feature models, such as latent factor models, are models where each observation is associated with a vector of latent features. A general problem is how to select the number/types of features, and related quantities. In Bayesian statistical machine learning, one seeks (nonparametric) models where one can learn such quantities in the presence of observed data. The Indian Buffet Process (IBP), devised by Griffiths and Ghahramani (2005), generates a (sparse) latent binary matrix with columns representing a potentially unbounded number of features and where each row corresponds to an individual or object. Its generative scheme is cast in terms of customers entering sequentially an Indian Buffet restaurant and selecting previously sampled dishes as well as new dishes. Dishes correspond to latent features shared by individuals. The IBP has been applied to a wide range of statistical problems. Recent works have demonstrated the utility of generalizations to nonbinary matrices. The purpose of this work is to describe a unified mechanism for construction, Bayesian analysis, and practical sampling of broad generalizations of the IBP that generate (sparse) matrices with general entries. An adaptation of the Poisson partition calculus is employed to handle the complexities, including combinatorial aspects, of these models. Our work reveals a spike and slab characterization, and also presents a general framework for multivariate extensions. We close by highlighting a multivariate IBP with condiments, and the role of a stable-Beta Dirichlet multivariate prior.

REFERENCES


MSC2010 subject classifications. Primary 60C05, 60G09; secondary 60G57, 60E99.

Key words and phrases. Bayesian statistical machine learning, Indian buffet process, nonparametric latent feature models, Poisson process calculus, spike and slab priors.


INFORMATION-REGRET COMPROMISE IN COVARIATE-ADAPTIVE TREATMENT ALLOCATION

BY ASYA METELKINA AND LUC PRONZATO

Université Côte d’Azur, CNRS, I3S

Covariate-adaptive treatment allocation is considered in the situation when a compromise must be made between information (about the dependency of the probability of success of each treatment upon influential covariates) and cost (in terms of number of subjects receiving the poorest treatment). Information is measured through a design criterion for parameter estimation, the cost is additive and is related to the success probabilities. Within the framework of approximate design theory, the determination of optimal allocations forms a compound design problem. We show that when the covariates are i.i.d. with a probability measure $\mu$, its solution possesses some similarities with the construction of optimal design measures bounded by $\mu$. We characterize optimal designs through an equivalence theorem and construct a covariate-adaptive sequential allocation strategy that converges to the optimum. Our new optimal designs can be used as benchmarks for other, more usual, allocation methods. A response-adaptive implementation is possible for practical applications with unknown model parameters. Several illustrative examples are provided.

REFERENCES


MSC2010 subject classifications. Primary 62K05; secondary 62P10.

Key words and phrases. Optimal design, treatment allocation, bounded design measure, equivalence theorem, multi-armed bandit problem.


SPARSE CCA: ADAPTIVE ESTIMATION AND COMPUTATIONAL BARRIERS

BY CHAO GAO¹, ZONGMING MA² AND HARRISON H. ZHOU¹

University of Chicago, University of Pennsylvania and Yale University

Canonical correlation analysis is a classical technique for exploring the relationship between two sets of variables. It has important applications in analyzing high dimensional datasets originated from genomics, imaging and other fields. This paper considers adaptive minimax and computationally tractable estimation of leading sparse canonical coefficient vectors in high dimensions. Under a Gaussian canonical pair model, we first establish separate minimax estimation rates for canonical coefficient vectors of each set of random variables under no structural assumption on marginal covariance matrices. Second, we propose a computationally feasible estimator to attain the optimal rates adaptively under an additional sample size condition. Finally, we show that a sample size condition of this kind is needed for any randomized polynomial-time estimator to be consistent, assuming hardness of certain instances of the planted clique detection problem. As a byproduct, we obtain the first computational lower bounds for sparse PCA under the Gaussian single spiked covariance model.

REFERENCES


MSC2010 subject classifications. Primary 62H12; secondary 62C20.
Key words and phrases. Convex programming, group-Lasso, minimax rates, computational complexity, planted clique, sparse CCA (SCCA), sparse PCA (SPCA).


OPTIMAL DESIGNS FOR DOSE RESPONSE CURVES WITH COMMON PARAMETERS

BY CHRYSTEL FELLER*, KIRSTEN SCHORNING†, HOLGER DETTE†, GEORGINA BERMANN* AND BJÖRN BORNKAMP*

Novartis Pharma AG* and Ruhr-Universität Bochum†

A common problem in Phase II clinical trials is the comparison of dose response curves corresponding to different treatment groups. If the effect of the dose level is described by parametric regression models and the treatments differ in the administration frequency (but not in the sort of drug), a reasonable assumption is that the regression models for the different treatments share common parameters.

This paper develops optimal design theory for the comparison of different regression models with common parameters. We derive upper bounds on the number of support points of admissible designs, and explicit expressions for $D$-optimal designs are derived for frequently used dose response models with a common location parameter. If the location and scale parameter in the different models coincide, minimally supported designs are determined and sufficient conditions for their optimality in the class of all designs derived. The results are illustrated in a dose-finding study comparing monthly and weekly administration.

REFERENCES


MSC2010 subject classifications. Primary 62K05; secondary 62F03.

Key words and phrases. Admissible design, different treatment groups, $D$-optimal design, Loewner ordering, nonlinear regression.


FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

BY WEIJIE SU¹, MAŁGORZATA BOGDAN² AND EMMANUEL CANDÈS³

University of Pennsylvania, University of Wroclaw and Stanford University

In regression settings where explanatory variables have very low correlations and there are relatively few effects, each of large magnitude, we expect the Lasso to find the important variables with few errors, if any. This paper shows that in a regime of linear sparsity—meaning that the fraction of variables with a nonvanishing effect tends to a constant, however small—this cannot really be the case, even when the design variables are stochastically independent. We demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the effect sizes are. We derive a sharp asymptotic trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the Lasso path. This trade-off states that if we ever want to achieve a type II error (false negative rate) under a critical value, then anywhere on the Lasso path the type I error (false positive rate) will need to exceed a given threshold so that we can never have both errors at a low level at the same time. Our analysis uses tools from approximate message passing (AMP) theory as well as novel elements to deal with a possibly adaptive selection of the Lasso regularizing parameter.

REFERENCES


MSC2010 subject classifications. Primary 62F03; secondary 62J07, 62J05.

Key words and phrases. Lasso, Lasso path, false discovery rate, false negative rate, power, approximate message passing (AMP), adaptive selection of parameters.


PHASE TRANSITIONS FOR HIGH DIMENSIONAL CLUSTERING AND RELATED PROBLEMS

BY JIASHUN JIN, ZHENG TRACY KE AND WANJIE WANG

Carnegie Mellon University, University of Chicago and University of Pennsylvania

Consider a two-class clustering problem where we observe $X_i = \ell_i \mu + Z_i$, $Z_i \sim \mathcal{N}(0, I_p)$, $1 \leq i \leq n$. The feature vector $\mu \in \mathbb{R}^p$ is unknown but is presumably sparse. The class labels $\ell_i \in \{-1, 1\}$ are also unknown and the main interest is to estimate them.

We are interested in the statistical limits. In the two-dimensional phase space calibrating the rarity and strengths of useful features, we find the precise demarcation for the Region of Impossibility and Region of Possibility. In the former, useful features are too rare/weak for successful clustering. In the latter, useful features are strong enough to allow successful clustering. The results are extended to the case of colored noise using Le Cam’s idea on comparison of experiments.

We also extend the study on statistical limits for clustering to that for signal recovery and that for global testing. We compare the statistical limits for three problems and expose some interesting insight.

We propose classical PCA and Important Features PCA (IF-PCA) for clustering. For a threshold $t > 0$, IF-PCA clusters by applying classical PCA to all columns of $X$ with an $L^2$-norm larger than $t$. We also propose two aggregation methods. For any parameter in the Region of Possibility, some of these methods yield successful clustering.

We discover a phase transition for IF-PCA. For any threshold $t > 0$, let $\xi(t)$ be the first left singular vector of the post-selection data matrix. The phase space partitions into two different regions. In one region, there is a $t$ such that $\cos(\xi(t), \ell) \to 1$ and IF-PCA yields successful clustering. In the other, $\cos(\xi(t), \ell) \leq c_0 < 1$ for all $t > 0$.

Our results require delicate analysis, especially on post-selection random matrix theory and on lower bound arguments.

REFERENCES


Key words and phrases. Clustering, comparison of experiments, feature selection, hypothesis testing, $L^1$-distance, lower bound, low-rank matrix recovery, phase transition.


BAYESIAN DETECTION OF IMAGE BOUNDARIES

BY MENG LI AND SUBHASHIS GHOSAL

Duke University and North Carolina State University

Detecting boundary of an image based on noisy observations is a fundamental problem of image processing and image segmentation. For a $d$-dimensional image ($d = 2, 3, \ldots$), the boundary can often be described by a closed smooth $(d - 1)$-dimensional manifold. In this paper, we propose a nonparametric Bayesian approach based on priors indexed by $S^{d-1}$, the unit sphere in $\mathbb{R}^d$. We derive optimal posterior contraction rates for Gaussian processes or finite random series priors using basis functions such as trigonometric polynomials for 2-dimensional images and spherical harmonics for 3-dimensional images. For 2-dimensional images, we show a rescaled squared exponential Gaussian process on $S^1$ achieves four goals of guaranteed geometric restriction, (nearly) minimax optimal rate adapting to the smoothness level, convenience for joint inference and computational efficiency. We conduct an extensive study of its reproducing kernel Hilbert space, which may be of interest by its own and can also be used in other contexts. Several new estimates on modified Bessel functions of the first kind are given. Simulations confirm excellent performance and robustness of the proposed method.

REFERENCES


Key words and phrases. Boundary detection, Gaussian process on sphere, image, posterior contraction rate, random series, squared exponential periodic kernel, Bayesian adaptation.


SPECTRUM ESTIMATION FROM SAMPLES

BY WEIHAO KONG AND GREGORY VALIANT

Stanford University

We consider the problem of approximating the set of eigenvalues of the covariance matrix of a multivariate distribution (equivalently, the problem of approximating the “population spectrum”), given access to samples drawn from the distribution. We consider this recovery problem in the regime where the sample size is comparable to, or even sublinear in the dimensionality of the distribution. First, we propose a theoretically optimal and computationally efficient algorithm for recovering the moments of the eigenvalues of the population covariance matrix. We then leverage this accurate moment recovery, via a Wasserstein distance argument, to accurately reconstruct the vector of eigenvalues. Together, this yields an eigenvalue reconstruction algorithm that is asymptotically consistent as the dimensionality of the distribution and sample size tend toward infinity, even in the sublinear sample regime where the ratio of the sample size to the dimensionality tends to zero. In addition to our theoretical results, we show that our approach performs well in practice for a broad range of distributions and sample sizes.

REFERENCES


MSC2010 subject classifications. 62H12, 62H10.

Key words and phrases. Spectrum estimation, eigenvalues of covariance matrices, sublinear sample size, method of moments, random matrix theory, high-dimensional inference.


ON THE CONTRACTION PROPERTIES OF SOME HIGH-DIMENSIONAL QUASI-POSTERIOR DISTRIBUTIONS

BY YVES A. ATCHADÉ

University of Michigan

We study the contraction properties of a quasi-posterior distribution \( \tilde{\Pi}_{n,d} \) obtained by combining a quasi-likelihood function and a sparsity inducing prior distribution on \( \mathbb{R}^d \), as both \( n \) (the sample size), and \( d \) (the dimension of the parameter) increase. We derive some general results that highlight a set of sufficient conditions under which \( \tilde{\Pi}_{n,d} \) puts increasingly high probability on sparse subsets of \( \mathbb{R}^d \), and contracts toward the true value of the parameter. We apply these results to the analysis of logistic regression models, and binary graphical models, in high-dimensional settings. For the logistic regression model, we show that for well-behaved design matrices, the posterior distribution contracts at the rate \( O(\sqrt{s^*_\star \log(d)/n}) \), where \( s^*_\star \) is the number of nonzero components of the parameter. For the binary graphical model, under some regularity conditions, we show that a quasi-posterior analog of the neighborhood selection of [Ann. Statist. 34 (2006) 1436–1462] contracts in the Frobenius norm at the rate \( O(\sqrt{(p + S) \log(p)/n}) \), where \( p \) is the number of nodes, and \( S \) the number of edges of the true graph.

REFERENCES


Key words and phrases. Quasi-Bayesian inference, high-dimensional inference, Bayesian asymptotics, logistic regression models, discrete graphical models.


NONASYMPTOTIC ANALYSIS OF SEMIPARAMETRIC REGRESSION MODELS WITH HIGH-DIMENSIONAL PARAMETRIC COEFFICIENTS

BY YING ZHU

Michigan State University

We consider a two-step projection based Lasso procedure for estimating a partially linear regression model where the number of coefficients in the linear component can exceed the sample size and these coefficients belong to the $l_q$-“balls” for $q \in [0, 1]$. Our theoretical results regarding the properties of the estimators are nonasymptotic. In particular, we establish a new nonasymptotic “oracle” result: Although the error of the nonparametric projection per se (with respect to the prediction norm) has the scaling $t_n$ in the first step, it only contributes a scaling $t_n^2$ in the $l_2$-error of the second-step estimator for the linear coefficients. This new “oracle” result holds for a large family of nonparametric least squares procedures and regularized nonparametric least squares procedures for the first-step estimation and the driver behind it lies in the projection strategy. We specialize our analysis to the estimation of a semiparametric sample selection model and provide a simple method with theoretical guarantees for choosing the regularization parameter in practice.

REFERENCES


MSC2010 subject classifications. Primary 62J02; secondary 62N01, 62N02, 62G08, 62J12.

Key words and phrases. High-dimensional statistics, Lasso, nonasymptotic analysis, partially linear models, sample selection.


