TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY, ISSN 1133-0686
Volume 31, number 1 (march 2022)

Abstract
Given k independent samples with finite but arbitrary dimension, this paper deals with the problem of testing for the equality of their distributions that can be continuous, discrete or mixed. In contrast to the classical setting where k is assumed to be fixed and the sample size from each population increases without bound, here k is assumed to be large and the size of each sample is either bounded or small in comparison with k. The asymptotic distribution of two test statistics is stated under the null hypothesis of the equality of the k distributions as well as under alternatives, which let us to study the asymptotic power of the resulting tests. Specifically, it is shown that both test statistics are asymptotically free distributed under the null hypothesis. The finite sample performance of the tests based on the asymptotic null distribution is studied via simulation. An application of the proposal to a real data set is included. The use of the proposed procedure for infinite dimensional data, as well as other possible extensions, are discussed.

Abstract
Mixtures of t factor analyzers (MtFA) have been well recognized as a prominent tool in modeling and clustering multivariate data contaminated with heterogeneity and outliers. In certain practical situations, however, data are likely to be censored such that the standard methodology becomes computationally complicated or even infeasible. This paper presents an extended framework of MtFA that can accommodate censored data, referred to as MtFAC in short. For maximum likelihood estimation, we construct an alternating expectation conditional maximization algorithm in which the E-step relies on the first-two moments of truncated multivariate-t distributions and CM-steps offer tractable solutions of updated estimators. Asymptotic standard errors of mixing proportions and component mean vectors are derived by means of missing information principle, or the so-called Louis' method. Several numerical experiments are conducted to examine the finite-sample properties of estimators and the ability of the proposed model to downweight the impact of censoring and outlying effects. Further, the efficacy and usefulness of the proposed method are also demonstrated by analyzing a real dataset with genuine censored observations.

Abstract
Penalized estimation can conduct variable selection and parameter estimation simultaneously. The general framework is to minimize a loss function subject to a penalty designed to generate sparse variable selection. The majorization–minimization (MM) algorithm is a computational scheme for stability and simplicity, and the MM algorithm has been widely applied in penalized estimation. Much of the previous work has focused on convex loss functions such as generalized linear models. When data are contaminated with outliers, robust loss functions can generate more reliable

estimates. Recent literature has witnessed a growing impact of nonconvex loss-based methods, which can generate robust estimation for data contaminated with outliers. This article investigates MM algorithm for penalized estimation, provides innovative optimality conditions and establishes convergence theory with both convex and nonconvex loss functions. With respect to applications, we focus on several nonconvex loss functions, which were formerly studied in machine learning for regression and classification problems. Performance of the proposed algorithms is evaluated on simulated and real data including cancer clinical status. Efficient implementations of the algorithms are available in the R package mpath in CRAN.

Goodness-of-fit test with a robustness feature P. 76-100

Jiming Jiang, Mahmoud Torabi

Abstract

We develop a method originally proposed by R. A. Fisher into a general procedure, called tailoring, for deriving goodness-of-fit tests that are guaranteed to have a $\chi 2$ asymptotic null distribution. The method has a robustness feature that it works correctly in testing a certain aspect of the model while some other aspect of the model may be misspecified. We apply the method to small area estimation. A connection, and difference, to the existing specification test is discussed. We evaluate performance of the tests both theoretically and empirically, and compare the performance with several existing methods. Our empirical results suggest that the proposed test is more accurate in size, and has either higher or similar power compared to the existing tests. The proposed test is also computationally less demanding than the specification test and other comparing methods. A real-data application is discussed.

Bayesian semiparametric modeling of response mechanism for nonignorable missing data P. 101-117

Shonosuke Sugasawa, Kosuke Morikawa, Keisuke Takahata

Abstract

Statistical inference with nonresponse is quite challenging, especially when the response mechanism is nonignorable. In this case, the validity of statistical inference depends on untestable correct specification of the response model. To avoid the misspecification, we propose semiparametric Bayesian estimation in which an outcome model is parametric, but the response model is semiparametric in that we do not assume any parametric form for the nonresponse variable. We adopt penalized spline methods to estimate the unknown function. We also consider a fully nonparametric approach to modeling the response mechanism by using radial basis function methods. Using Pólya–gamma data augmentation, we developed an efficient posterior computation algorithm via Gibbs sampling in which most full conditional distributions can be obtained in familiar forms. The performance of the proposed method is demonstrated in simulation studies and an application to longitudinal data.

Robust parametric inference for finite Markov chains P. 118–147

Abhik Ghosh

Abstract

We consider the problem of statistical inference in a parametric finite Markov chain model and develop a robust estimator of the parameters defining the transition probabilities via minimization of a suitable (empirical) version of the popular density power divergence. Based on a long sequence of observations from a first-order stationary Markov chain, we have defined the minimum density power divergence estimator (MDPDE) of the underlying parameter and rigorously derived its asymptotic and robustness properties under appropriate conditions. Performance of the MDPDEs is illustrated theoretically as well as empirically for some common examples of finite Markov chain models. Its applications in robust testing of statistical hypotheses are also discussed along with (parametric) comparison of two Markov chain sequences. Several directions for extending the MDPDE and related inference are also briefly discussed for multiple sequences of Markov chains, higher order Markov chains and non-stationary Markov chains with time-dependent transition probabilities. Finally, our proposal is applied to analyze corporate credit rating migration data of

three international markets.

Where to find needles in a haystack? P. 148–174

Zhigen Zhao

Abstract

In many existing methods of multiple comparison, one starts with either Fisher's *p* value or the local fdr. One commonly used *p* value, defined as the tail probability exceeding the observed test statistic under the null distribution, fails to use information from the distribution under the alternative hypothesis. The targeted region of signals could be wrong when the likelihood ratio is not monotone. The oracle local fdr based approaches could be optimal because they use the probability density functions of the test statistic under both the null and alternative hypotheses. However, the data-driven version could be problematic because of the difficulty and challenge of probability density function estimation. In this paper, we propose a new method, Cdf and Local fdr Assisted multiple Testing method (CLAT), which is optimal for cases when the *p* value based methods are optimal and for some other cases when *p* value based methods are not. Additionally, CLAT only relies on the empirical distribution function which quickly converges to the oracle one. Both the simulations and real data analysis demonstrate the superior performance of the CLAT method. Furthermore, the computation is instantaneous based on a novel algorithm and is scalable to large data sets.

Spatial Cox processes in an infinite-dimensional framework P. 175-203

María P. Frías, Antoni Torres-Signes, Jorge Mateu

Abstract

We introduce a new class of spatial Cox processes driven by a Hilbert-valued random log-intensity. We adopt a parametric framework in the spectral domain, to estimate its spatial functional correlation structure. Specifically, we consider a spectral functional, approach based on the periodogram operator, inspired on Whittle estimation methodology. Strong consistency of the parametric estimator is proved in the linear case. We illustrate this property in a simulation study under a Gaussian first-order Spatial Autoregressive Hilbertian scenario for the log-intensity model. Our method is applied to the spatial functional prediction of respiratory disease mortality in the Spanish Iberian Peninsula, in the period 1980–2015.

A measurement error Rao–Yu model for regional prevalence estimation over time using uncertain data obtained from dependent survey estimates P. 204-234

Jan Pablo Burgard, Joscha Krause, Domingo Morales

Abstract

The assessment of prevalence on regional levels is an important element of public health reporting. Since regional prevalence is rarely collected in registers, corresponding figures are often estimated via small area estimation using suitable health data. However, such data are frequently subject to uncertainty as values have been estimated from surveys. In that case, the method for prevalence estimation must explicitly account for data uncertainty to allow for reliable results. This can be achieved via measurement error models that introduce distribution assumptions on the noisy data. However, these methods usually require target and explanatory variable errors to be independent. This does not hold when data for both have been estimated from the same survey, which is sometimes the case in official statistics. If not accounted for, prevalence estimates can be severely biased. We propose a new measurement error model for regional prevalence estimation that is suitable for settings where target and explanatory variable errors are dependent. We derive empirical best predictors and demonstrate mean-squared error estimation. A maximum likelihood approach for model parameter estimation is presented. Simulation experiments are conducted to prove the effectiveness of the method. An application to regional hypertension prevalence estimation in Germany is provided.

A new inferential approach for response-adaptive clinical trials: the variance- P. 235-254

stabilized bootstrap

Alessandro Baldi Antognini, Marco Novelli, Maroussa Zagoraiou

Abstract

This paper discusses disadvantages and limitations of the available inferential approaches in sequential clinical trials for treatment comparisons managed via response-adaptive randomization. Then, we propose an inferential methodology for response-adaptive designs which, by exploiting a variance stabilizing transformation into a bootstrap framework, is able to overcome the above-mentioned drawbacks, regardless of the chosen allocation procedure as well as the desired target. We derive the theoretical properties of the suggested proposal, showing its superiority with respect to likelihood, randomization and design-based inferential approaches. Several illustrative examples and simulation studies are provided in order to confirm the relevance of our results.

Sparse Laplacian Shrinkage with the Graphical Lasso Estimator for Regression Problems P. 255-277

Siwei Xia, Yuehan Yang, Hu Yang

Abstract

This paper considers a high-dimensional linear regression problem where there are complex correlation structures among predictors. We propose a graph-constrained regularization procedure, named Sparse Laplacian Shrinkage with the Graphical Lasso Estimator (SLS-GLE). The procedure uses the estimated precision matrix to describe the specific information on the conditional dependence pattern among predictors, and encourages both sparsity on the regression model and the graphical model. We introduce the Laplacian quadratic penalty adopting the graph information, and give detailed discussions on the advantages of using the precision matrix to construct the Laplacian matrix. Theoretical properties and numerical comparisons are presented to show that the proposed method improves both model interpretability and accuracy of estimation. We also apply this method to a financial problem and prove that the proposed procedure is successful in assets selection.

Bayesian and frequentist evidence in one-sided hypothesis testing P. 278-297

Elías Moreno, Carmen Martínez

Abstract

In one-sided testing, Bayesians and frequentists differ on whether or not there is discrepancy between the inference based on the posterior model probability and that based on the p value. We add some arguments to this debate analyzing the discrepancy for moderate and large sample sizes. For small and moderate samples sizes, the discrepancy is measured by the probability of disagreement. Examples of the discrepancy on some basic sampling models indicate the somewhat unexpected result that the probability of disagreement is larger when sampling from models in the alternative hypothesis that are not located at the boundary of the hypotheses. For large sample sizes, we prove that the Bayesian one-sided testing is, under mild conditions, consistent, a property that is not shared by the frequentist procedure. Further, the rate of convergence is $O(enA)$, where A is a constant that depends on the model from which we are sampling. Consistency is also proved for an extension to multiple hypotheses.