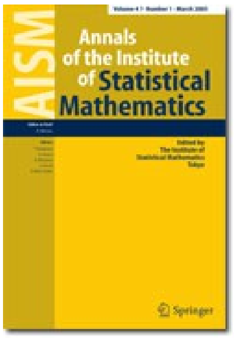

Annals of the Institute of Statistical Mathematics, ISSN 0020-3157
Volume 67, number 2 (april 2015)

An empirical estimator for the sparsity of a large covariance matrix under multivariate normal assumptions

P. 211-227

Binyan Jiang

Abstract

Large covariance or correlation matrix is frequently assumed to be sparse in that a number of the off-diagonal elements of the matrix are zero. This paper focuses on estimating the sparsity of a large population covariance matrix using a sample correlation matrix under multivariate normal assumptions. We show that sparsity of a population covariance matrix can be well estimated by thresholding the sample correlation matrix. We then propose an empirical estimator for the sparsity and show that it is closely related to the thresholding methods. Upper bounds for the estimation error of the empirical estimator are given under mild conditions. Simulation shows that the empirical estimator can have smaller mean absolute errors than its main competitors. Furthermore, when the dimension of the covariance matrix is very large, we propose a generalized empirical estimator using simple random sampling. It is shown that the generalized empirical estimator can still estimate the sparsity well while the computation complexity can be greatly reduced.

Model checking for parametric regressions with response missing at random

P. 229-259

Xu Guo, Wangli Xu, Lixing Zhu

Abstract

This paper aims at investigating model checking for parametric models with response missing at random which is a more general missing mechanism than missing completely at random. Different from existing approaches, two tests have normal distributions as the limiting null distributions no matter whether the inverse probability weight is estimated parametrically or nonparametrically. Thus, TeX values can be easily determined. This observation shows that slow convergence rate of nonparametric estimation does not have significant effect on the asymptotic behaviors of the tests although it may have impact in finite sample scenarios. The tests can detect the alternatives distinct from the null hypothesis at a nonparametric rate which is an optimal rate for locally smoothing-based methods in this area. Simulation study is carried out to examine the performance of the tests. The tests are also applied to analyze a data set on monozygotic twins for illustration.

Minimaxity in estimation of restricted and non-restricted scale parameter matrices

P. 261-285

Hisayuki Tsukuma, Tatsuya Kubokawa

Abstract

In estimation of the normal covariance matrix, finding a least favorable sequence of prior distributions has been an open question for a long time. This paper addresses the classical problem and accomplishes the specification of such a sequence, which establishes minimaxity of the best equivariant estimator. This result is extended to the estimation of scale parameter matrix in an elliptically contoured distribution model. The methodology based on a least favorable sequence of prior distributions is applied to both restricted and non-restricted cases of parameters, and we give some examples which show minimaxity of the best equivariant estimators under restrictions of scale parameter matrix.

Extended Bayesian information criterion in the Cox model with a high-dimensional feature space

P. 287-311

Shan Luo, Jinfeng Xu, Zehua Chen

Abstract

Variable selection in the Cox proportional hazards model (the Cox model) has manifested its importance in many microarray genetic studies. However, theoretical results on the procedures of variable selection in the Cox model with a high-dimensional feature space are rare because of its complicated data structure. In this paper, we consider the extended Bayesian information criterion (EBIC) for variable selection in the Cox model and establish its selection consistency in the situation of high-dimensional feature space. The EBIC is adopted to select the best model from a model sequence generated from the SIS-ALasso procedure. Simulation studies and real data analysis are carried out to demonstrate the merits of the EBIC.

A normal hierarchical model and minimum contrast estimation for random intervals

P. 313-333

Yan Sun, Dan Ralescu

Abstract

Many statistical data are imprecise due to factors such as measurement errors, computation errors, and lack of information. In such cases, data are better represented by intervals rather than by single numbers. Existing methods for analyzing interval-valued data include regressions in the metric space of intervals and symbolic data analysis, the latter being proposed in a more general setting. However, there has been a lack of literature on the parametric modeling and distribution-based inferences for interval-valued data. In an attempt to fill this gap, we extend the concept of normality for random sets by Lyashenko and propose a Normal hierarchical model for random intervals. In addition, we develop a minimum contrast estimator (MCE) for the model parameters, which is both consistent and asymptotically normal. Simulation studies support our theoretical findings and show very promising results. Finally, we successfully apply our model and MCE to a real data set.

Strong consistency of factorial TeX -means clustering

P. 335-357

Yoshikazu Terada

Abstract

Factorial TeX -means (FKM) clustering is a method for clustering objects in a low-dimensional subspace. The advantage of this method is that the partition of objects and the low-dimensional subspace reflecting the cluster structure are obtained, simultaneously. In some cases that reduced TeX -means (RKM) clustering does not work well, FKM clustering can discover the cluster structure underlying a lower dimensional subspace. Conditions that ensure the almost sure convergence of the estimator of FKM clustering as the sample size increases unboundedly are derived. The result is proved for a more general model including FKM clustering. Moreover, it is also shown that there exist some cases in which RKM clustering becomes equivalent to FKM clustering as the sample size goes to infinity.

On generalized expectation-based estimation of a population spectral distribution from high-dimensional data

P. 359-373

Weiming Li, Jianfeng Yao

Abstract

This paper discusses the problem of estimating the population spectral distribution from high-dimensional data. We present a general estimation procedure that covers situations where the moments of this distribution fail to identify the model parameters. The main idea is to use generalized functional expectations as a substitute for the moments. Beyond the consistency, we also prove a central limit theorem for the proposed estimator. Simulation experiments illustrate the implementation of the estimation procedure. An application to the analysis of the eigenvalues of the sample correlation matrix of S&P 500 daily stock returns is proposed.

Quantile regression and variable selection of partial linear single-index model

P. 375-409

Yazhao Lv, Riquan Zhang, Weihua Zhao...

Abstract

Partial linear single-index model (PLSIM) is a flexible and applicable model when investigating the underlying relationship between the response and the multivariate covariates. Most previous studies on PLSIM concentrated on mean regression, based on least square or likelihood approach. In contrast to this method, in this paper, we propose minimizing average check loss estimation (MACLE) procedure to conduct quantile regression of PLSIM. We construct an initial consistent quantile regression estimator of the parametric part base multi-dimensional kernels, and further promote the estimation efficiency to the optimal rate. We discuss the optimal bandwidth selection method and establish the asymptotic normality of the proposed MACLE estimators. Furthermore, we consider an adaptive lasso penalized variable selection method and establish its oracle property. Simulation studies with various distributed error and a real data analysis are conducted to show the promise of our proposed methods.