Biometrika, ISSN 0006-3444
Volume 102, number 1 (march 2015)

Abstract
A characteristic feature of functional data is the presence of phase variability in addition to amplitude variability. Existing functional regression methods do not handle time variability in an explicit and efficient way. In this paper we introduce a functional regression method that incorporates time warping as an intrinsic part of the model. The method achieves good predictive power in a parsimonious way and allows unified statistical inference about phase and amplitude components. The asymptotic distribution of the estimators is derived and their finite-sample properties are studied by simulation. An application involving ground-level ozone trajectories is presented.

Abstract
Both varying-coefficient and additive models have been studied extensively in the literature as extensions to linear models. They have also been extended to deal with functional response data. However, existing extensions are still not flexible enough to reflect the functional nature of the responses. In this paper, we extend varying-coefficient and additive models to obtain a much more flexible model and propose a simple algorithm to estimate its nonparametric additive and varying-coefficient components. We establish the asymptotic properties of each component function. We demonstrate the applicability of the new model through analysis of traffic data.

Abstract
Linear discriminant analysis has been widely used to characterize or separate multiple classes via linear combinations of features. However, the high dimensionality of features from modern biological experiments defies traditional discriminant analysis techniques. Possible interfeature correlations present additional challenges and are often underused in modelling. In this paper, by incorporating possible interfeature correlations, we propose a covariance-enhanced discriminant analysis method that simultaneously and consistently selects informative features and identifies the corresponding discriminable classes. Under mild regularity conditions, we show that the method can achieve consistent parameter estimation and model selection, and can attain an asymptotically optimal misclassification rate. Extensive simulations have verified

Abstract
We consider the problem of estimating the parameters in a pairwise graphical model in which the distribution of each node,

conditioned on the others, may have a different exponential family form. We identify restrictions on the parameter space required for the existence of a well-defined joint density, and establish the consistency of the neighbourhood selection approach for graph reconstruction in high dimensions when the true underlying graph is sparse. Motivated by our theoretical results, we investigate the selection of edges between nodes whose conditional distributions take different parametric forms, and show that efficiency can be gained if edge estimates obtained from the regressions of particular nodes are used to reconstruct the graph. These results are illustrated with examples of Gaussian, Bernoulli, Poisson and exponential distributions. Our theoretical findings are corroborated by evidence from simulation studies.

Abstract
To accommodate the heterogeneity that is often present in ultrahigh-dimensional data, we propose a conditional quantile screening method, which enables us to select features that contribute to the conditional quantile of the response given the covariates. The method can naturally handle censored data by incorporating a weighting scheme through redistribution of the mass to the right; moreover, it is invariant to monotone transformation of the response and requires substantially weaker conditions than do alternative methods. We establish sure independent screening properties for both the complete and the censored response cases. We also conduct simulations to evaluate the finite-sample performance of the proposed method, and compare it with existing approaches.

Abstract
We develop uniformly valid confidence regions for regression coefficients in a high-dimensional sparse median regression model with homoscedastic errors. Our methods are based on a moment equation that is immunized against nonregular estimation of the nuisance part of the median regression function by using Neyman's orthogonalization. We establish that the resulting instrumental median regression estimator of a target regression coefficient is asymptotically normally distributed uniformly with respect to the underlying sparse model and is semiparametrically efficient. We also generalize our method to a general nonsmooth Z-estimation framework where the number of target parameters is possibly much larger than the sample size. We extend Huber's results on asymptotic normality to this setting, demonstrating uniform asymptotic normality of the proposed estimators over rectangles, constructing simultaneous confidence bands on all of the target parameters, and establishing asymptotic validity of the bands uniformly over underlying approximately sparse models.

Abstract
Sufficient dimension reduction is a useful tool for studying the dependence between a response and a multi-dimensional predictor. In this article, a new formulation is proposed that is based on the Hellinger integral of order two, introduced as a natural measure of the regression information contained in the predictor subspace. The response may be either continuous or discrete. We establish links between local and global central subspaces, and propose an efficient local estimation algorithm. Simulations and an application show that our method compares favourably with existing approaches.

J. Shao - J. Zhang

Abstract

We consider a linear mixed-effects model in which the response panel vector has missing components and the missing data mechanism depends on observed data as well as missing responses through unobserved random effects. Using a transformation of the data that eliminates the random effects, we derive asymptotically unbiased and normally distributed estimators of certain model parameters. Estimators of model parameters that cannot be estimated using the transformed data are also constructed, and their asymptotic unbiasedness and normality are established. Simulation results are presented to examine the finite sample performance of the proposed estimators and a real data example is discussed.

Moment-type estimators for the proportional likelihood ratio model with longitudinal data P. 121-134

Xiaodong Luo - Wei Yann Tsai

Abstract

Luo & Tsai, *Biometrika* 99, 211–22, 2012, proposed a proportional likelihood ratio model and discussed a maximum likelihood method for its parameter estimation. In this paper, we use this model as the marginal distribution to analyse longitudinal data, where the maximum likelihood method is not directly applicable because the joint distribution is not fully specified. We propose a moment-type method that is an extension of the generalized estimating equation method. The resulting estimators are consistent, asymptotically normal and perform well in our simulation study.

An extended hazard model with longitudinal covariates P. 135-150

Y. K. Tseng - Y. R. Su - M. Mao - J. L. Wang

Abstract

In clinical trials and other medical studies, it has become increasingly common to observe simultaneously an event time of interest and longitudinal covariates. In the literature, joint modelling approaches have been employed to analyse both survival and longitudinal processes and to investigate their association. However, these approaches focus mostly on developing adaptive and flexible longitudinal processes based on a prespecified survival model, most commonly the Cox proportional hazards model. In this paper, we propose a general class of semiparametric hazard regression models, referred to as the extended hazard model, for the survival component. This class includes two popular survival models, the Cox proportional hazards model and the accelerated failure time model, as special cases. The proposed model is flexible for modelling event data, and its nested structure facilitates model selection for the survival component through likelihood ratio tests. A pseudo joint likelihood approach is proposed for estimating the unknown parameters and components via a Monte Carlo em algorithm. Asymptotic theory for the estimators is developed together with theory for the semiparametric likelihood ratio tests. The performance of the procedure is demonstrated through simulation studies. A case study featuring data from a Taiwanese HIV/AIDS cohort study further illustrates the usefulness of the extended hazard model.

Doubly robust learning for estimating individualized treatment with censored data P. 151-168

Y. Q. Zhao - D. Zeng - E. B. Laber - R. Song - M. Yuan - M. R. Kosorok

Abstract

Individualized treatment rules recommend treatments based on individual patient characteristics in order to maximize clinical benefit. When the clinical outcome of interest is survival time, estimation is often complicated by censoring. We develop nonparametric methods for estimating an optimal individualized treatment rule in the presence of censored data. To adjust for censoring, we propose a doubly robust estimator which requires correct specification of either the censoring model or survival model, but not both; the method is shown to be Fisher consistent when either model is correct. Furthermore, we establish the

convergence rate of the expected survival under the estimated optimal individualized treatment rule to the expected survival under the optimal individualized treatment rule. We illustrate the proposed methods using simulation study and data from a Phase III clinical trial on non-small cell lung cancer.

Using covariate-specific disease prevalence information to increase the power of case-control studies P. 169-180

Jing Qin - Han Zhang - Pengfei Li - Demetrius Albanes - Kai Yu

Abstract

Public registration databases and large cohort studies provide vital information on disease prevalence at various levels of a risk factor. This auxiliary information can be helpful in conducting statistical inference in a new study. We aim to develop a statistical procedure that improves the efficiency of the logistic regression model for a case-control study by utilizing auxiliary information on covariate-specific disease prevalence via a series of unbiased estimating equations. We adopt empirical likelihood for statistical inference, and demonstrate its advantages through simulation and an application.

A tractable and interpretable four-parameter family of unimodal distributions on the circle P. 181-190

Shogo Kato - M. C. Jones

Abstract

This article presents a class of four-parameter distributions for circular data that are unimodal, possess simple characteristic and density functions and a tractable distribution function, can be interpretably parameterized directly in terms of their trigonometric moments, afford a very wide range of skewness and kurtosis, envelop numerous interesting submodels including the wrapped Cauchy and cardioid distributions, allow straightforward parameter estimation by both method of moments and maximum likelihood, and are closed under convolution. This class of distributions exhibits the widest range of attractive properties yet available while retaining unimodality.

Adaptive randomized trial designs that cannot be dominated by any standard design at the same total sample size P. 191-202

Michael Rosenblum

Abstract

Prior work has shown that the power of adaptive designs with rules for modifying the sample size can always be matched or improved by suitably chosen, standard, group sequential designs. A natural question is whether analogous results hold for other types of adaptive designs. We focus on adaptive enrichment designs, which involve preplanned rules for modifying enrollment criteria based on accrued data in a randomized trial. Such designs often involve multiple hypotheses, e.g., one for the total population and one for a predefined subpopulation, such as those with high disease severity at baseline. We fix the total sample size, and consider overall power, defined as the probability of rejecting at least one false null hypothesis. We present adaptive enrichment designs whose overall power at two alternatives cannot simultaneously be matched by any standard design. In some scenarios there is a substantial gap between the overall power achieved by these adaptive designs and that of any standard design. We also prove that such gains in overall power come at a cost. To attain overall power above what is achievable by certain standard designs, it is necessary to increase power to reject some hypotheses and reduce power to reject others. We demonstrate that adaptive enrichment designs allow certain power trade-offs that are not available when restricting to standard designs.

Double-bootstrap methods that use a single double-bootstrap simulation P. 203-214

Jinyuan Chang - Peter Hall

Abstract

We show that, when the double bootstrap is used to improve performance of bootstrap methods for bias correction, techniques based on using a single double-bootstrap sample for each single-bootstrap sample can produce third-order accuracy for much less computational expense than is required by conventional double-bootstrap methods. However, this improved level of performance is not available for the single double-bootstrap methods that have been suggested to construct confidence intervals or distribution estimators.

Multivariate max-stable spatial processes P. 215-230

Marc G. Genton - Simone A. Padoan - Huiyan Sang

Abstract

Max-stable processes allow the spatial dependence of extremes to be modelled and quantified, so they are widely adopted in applications. For a better understanding of extremes, it may be useful to study several variables simultaneously. To this end, we study the maxima of independent replicates of multivariate processes, both in the Gaussian and Student-t cases. We define a Poisson process construction and introduce multivariate versions of the Smith Gaussian extreme-value, the Schlather extremal-Gaussian and extremal-t, and the Brown–Resnick models. We develop inference for the models based on composite likelihoods. We present results of Monte Carlo simulations and an application to daily maximum wind speed and wind gust.

Generalized Ewens–Pitman model for Bayesian clustering P. 231-238

Harry Crane

Abstract

We propose a Bayesian method for clustering from discrete data structures that commonly arise in genetics and other applications. This method is equivariant with respect to relabelling units; unsampled units do not interfere with sampled data; and missing data do not hinder inference. Cluster inference using the posterior mode performs well on simulated and real datasets, and the posterior predictive distribution enables supervised learning based on a partial clustering of the sample.

A Wilcoxon–Mann–Whitney-type test for infinite-dimensional data P. 239-246

Anirvan Chakraborty - Probal Chaudhuri

Abstract

The Wilcoxon–Mann–Whitney test is a robust competitor of the t test in the univariate setting. For finite-dimensional multivariate non-Gaussian data, several extensions of the Wilcoxon–Mann–Whitney test have been shown to outperform Hotelling's T^2 test. In this paper, we study a Wilcoxon–Mann–Whitney-type test based on spatial ranks in infinite-dimensional spaces, we investigate its asymptotic properties and compare it with several existing tests. The proposed test is shown to be robust with respect to outliers and to have better power than some competitors for certain distributions with heavy tails. We study its performance using real and simulated data.