Biometrika, ISSN 0006-3444
Volume 102, number 3 (september 2015)

Abstract
Individualized treatment rules recommend treatments on the basis of individual patient characteristics. A high-quality treatment rule can produce better patient outcomes, lower costs and less treatment burden. If a treatment rule learned from data is to be used to inform clinical practice or provide scientific insight, it is crucial that it be interpretable; clinicians may be unwilling to implement models they do not understand, and black-box models may not be useful for guiding future research. The canonical example of an interpretable prediction model is a decision tree. We propose a method for estimating an optimal individualized treatment rule within the class of rules that are representable as decision trees. The class of rules we consider is interpretable but expressive. A novel feature of this problem is that the learning task is unsupervised, as the optimal treatment for each patient is unknown and must be estimated. The proposed method applies to both categorical and continuous treatments and produces favourable marginal mean outcomes in simulation experiments. We illustrate it using data from a study of major depressive disorder.

Abstract
With the discovery of an increasing number of causal genes for complex human disorders, it is crucial to assess the genetic risk of disease onset for individuals who are carriers of these causal mutations and to compare the distribution of the age-at-onset for such individuals with the distribution for noncarriers. In many genetic epidemiological studies that aim to estimate causal gene effect on disease, the age-at-onset of disease is subject to censoring. In addition, the mutation carrier or noncarrier status of some individuals may be unknown, due to the high cost of in-person ascertainment by collecting DNA samples or because of the death of older individuals. Instead, the probability of such individuals' mutation status can be obtained from various other sources. When mutation status is missing, the available data take the form of censored mixture data. Recently, various methods have been proposed for risk estimation using such data, but none is efficient for estimating a nonparametric distribution. We propose a fully efficient sieve maximum likelihood estimation method, in which we estimate the logarithm of the hazard ratio between genetic mutation groups using B-splines, while applying nonparametric maximum likelihood estimation to the reference baseline hazard function. Our estimator can be calculated via an expectation-maximization algorithm which is much faster than existing methods. We show that our estimator is consistent and semiparametrically efficient and establish its asymptotic distribution. Simulation studies demonstrate the superior performance of the proposed method, which is used to estimate the distribution of the age-at-onset of Parkinson's disease for carriers of mutations in the leucine-rich repeat kinase 2, LRRK2, gene.

Covariance-based analyses of biological pathways P. 533-544

P. Danaher - D. Paul - P. Wang

Abstract

The use of high-throughput data to study the changing behaviour of biological pathways has focused mainly on examining the changes in the means of pathway genes. In this paper, we propose instead to test for changes in the co-regulated and unregulated variability of pathway genes. We assume that the eigenvalues of previously defined pathways capture biologically relevant quantities, and we develop a test for biologically meaningful changes in the eigenvalues between classes. This test reflects important and often ignored aspects of pathway behaviour and provides a useful complement to traditional pathway analyses.

Diagnostic studies in sufficient dimension reduction P. 545-558

Xin Chen - R. Dennis Cook - Changliang Zou

Abstract

Sufficient dimension reduction in regression aims to reduce the predictor dimension by replacing the original predictors with some set of linear combinations of them without loss of information. Numerous dimension reduction methods have been developed based on this paradigm. However, little effort has been devoted to diagnostic studies within the context of dimension reduction. In this paper we introduce methods to check goodness-of-fit for a given dimension reduction subspace. The key idea is to extend the so-called distance correlation to measure the conditional dependence relationship between the covariates and the response given a reduction subspace. Our methods require minimal assumptions, which are usually much less restrictive than the conditions needed to justify the original methods. Asymptotic properties of the test statistic are studied. Numerical examples demonstrate the effectiveness of the proposed approach.

Robust estimation under heavy contamination using unnormalized models P. 559-572

Takafumi Kanamori - Hironori Fujisawa

Abstract

Contamination caused by outliers is inevitable in data analysis, and robust statistical methods are often needed. In this paper we develop a new approach for robust data analysis on the basis of scoring rules. A scoring rule is a discrepancy measure to assess the quality of probabilistic forecasts. We propose a simple method of estimating not only parameters in the statistical model but also the contamination ratio, i.e., the ratio of outliers. The outliers are detected based on the estimated contamination ratio. For this purpose, we use scoring rules with extended statistical models called unnormalized models. Regression problems are also considered. We study complex heterogeneous contamination wherein the contamination ratio in a response variable may depend on covariate variables, and propose a simple method to estimate a robust regression function and expected contamination ratio. Simulation studies demonstrate the effectiveness of our method.

A cautionary note on robust covariance plug-in methods P. 573-588

Klaus Nordhausen - David E. Tyler

Abstract

The sample covariance matrix, which is well known to be highly nonrobust, plays a central role in many classical multivariate statistical methods. A popular way of making such multivariate methods more robust is to replace the sample covariance matrix with some robust scatter matrix. The aim of this paper is to point out that multivariate methods often require that certain properties of the covariance matrix hold also for the robust scatter matrix in order for the corresponding robust plug-in method to be a valid approach, but that not all scatter matrices possess the desired properties. Plug-in methods for independent components analysis, observational regression and graphical modelling are considered in more detail. For each case, it is shown that replacing the sample covariance matrix with a symmetrized robust scatter matrix yields a valid robust multivariate procedure.

Outlier detection for high-dimensional data P. 589-599

Kwangil Ro - Changliang Zou - Zhaojun Wang - Guosheng Yin

Abstract

Outlier detection is an integral component of statistical modelling and estimation. For high-dimensional data, classical methods based on the Mahalanobis distance are usually not applicable. We propose an outlier detection procedure that replaces the classical minimum covariance determinant estimator with a high-breakdown minimum diagonal product estimator. The cut-off value is obtained from the asymptotic distribution of the distance, which enables us to control the Type I error and deliver robust outlier detection. Simulation studies show that the proposed method behaves well for high-dimensional data.

Bayesian sensitivity analysis with the Fisher–Rao metric P. 601-616

Sebastian Kurtek - Karthik Bharath

Abstract

We propose a geometric framework to assess sensitivity of Bayesian procedures to modelling assumptions based on the nonparametric Fisher–Rao metric. While the framework is general, the focus of this article is on assessing local and global robustness in Bayesian procedures with respect to perturbations of the likelihood and prior, and on the identification of influential observations. The approach is based on a square-root representation of densities, which enables analytical computation of geodesic paths and distances, facilitating the definition of naturally calibrated local and global discrepancy measures. An important feature of our approach is the definition of a geometric ϵ-contamination class of sampling distributions and priors via intrinsic analysis on the space of probability density functions. We demonstrate the applicability of our framework to generalized mixed-effects models and to directional and shape data.

Nonparametric Bayesian testing for monotonicity P. 617-630

J. G. Scott - T. S. Shively - S. G. Walker

Abstract

This paper adopts a nonparametric Bayesian approach to testing whether a function is monotone. Two new families of tests are constructed. The first uses constrained smoothing splines with a hierarchical stochastic-process prior that explicitly controls the prior probability of monotonicity. The second uses regression splines together with two proposals for the prior over the regression coefficients. Via simulation, the finite-sample performance of the tests is shown to improve upon existing frequentist and Bayesian methods. The asymptotic properties of the Bayes factor for comparing monotone versus nonmonotone regression functions in a Gaussian model are also studied. Our results significantly extend those currently available, which chiefly focus on determining the dimension of a parametric linear model.

Efficient computation of smoothing splines via adaptive basis sampling P. 631-645

Ping Ma - Jianhua Z. Huang - Nan Zhang

Abstract

Smoothing splines provide flexible nonparametric regression estimators. However, the high computational cost of smoothing splines for large datasets has hindered their wide application. In this article, we develop a new method, named adaptive basis sampling, for efficient computation of smoothing splines in super-large samples. Except for the univariate case where the Reinsch algorithm is applicable, a smoothing spline for a regression problem with sample size n can be expressed as a linear combination of n basis functions and its computational complexity is generally $O(n^3)$. We achieve a more scalable computation in the multivariate case by evaluating the smoothing spline using a smaller set of basis functions, obtained by an adaptive sampling scheme that uses values of the response variable. Our asymptotic analysis shows that smoothing splines computed via adaptive basis sampling converge to the true function at the same rate as full basis smoothing splines. Using simulation studies and a large-scale deep earth core-mantle boundary imaging study, we show that the proposed method outperforms a sampling method that does not

use the values of response variables.

Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation P. 647-659

M. Ghosh - T. Kubokawa - Y. Kawakubo

Abstract

The paper develops hierarchical empirical Bayes and benchmarked hierarchical empirical Bayes estimators of positive small area means under multiplicative models. The usual benchmarking requirement is that the small area estimates, when aggregated, should equal the direct estimates for the larger geographical areas. However, while estimating positive small area parameters, the conventional squared error or weighted squared error loss subject to the usual benchmark constraint may not produce positive estimators, so it is necessary to seek other loss functions. We consider a multiplicative model for the original data for estimating positive small area means, and suggest a variant of the Kullback–Leibler divergence as a loss function. The prediction errors of the suggested hierarchical empirical Bayes estimators are investigated asymptotically, and their second-order unbiased estimators are provided. Bootstrapped estimators of these prediction errors for both hierarchical empirical Bayes and benchmarked hierarchical empirical Bayes estimators are also given. The performance of the suggested procedures is investigated through simulation as well as with an example.

Entropy testing for nonlinear serial dependence in time series P. 661-675

Simone Giannerini - Esfandiar Maasoumi - Estela Bee Dagum

Abstract

We propose tests for nonlinear serial dependence in time series under the null hypothesis of general linear dependence, in contrast to the more widely studied null hypothesis of independence. The approach is based on combining an entropy dependence metric, which possesses many desirable properties and is used as a test statistic, with a suitable extension of surrogate data methods, a class of Monte Carlo distribution-free tests for nonlinearity, and a smoothed sieve bootstrap scheme. We show how, in the same way as the autocorrelation function is used for linear models, our tests can in principle be employed to detect the lags at which a significant nonlinear relationship is present. We prove the asymptotic validity of the proposed procedures and the corresponding inferences. The small-sample performance of the tests in terms of power and size is assessed through a simulation study. Applications to real datasets of different kinds are also presented.

Designs for generalized linear models with random block effects via information matrix approximations P. 677-693

T. W. Waite - D. C. Woods

Abstract

The selection of optimal designs for generalized linear mixed models is complicated by the fact that the Fisher information matrix, on which most optimality criteria depend, is computationally expensive to evaluate. We provide two novel approximations that reduce the computational cost of evaluating the information matrix by complete enumeration of response outcomes, or Monte Carlo approximations thereof: an asymptotic approximation that is accurate when there is strong dependence between observations in the same block; and an approximation via kriging interpolators. For logistic random intercept models, we show how interpolation can be especially effective for finding pseudo-Bayesian designs that incorporate uncertainty in the values of the model parameters. The new results are used to evaluate the efficiency, for estimating conditional models, of optimal designs from closed-form approximations to the information matrix derived from marginal models. Correcting for the marginal attenuation of parameters in binary-response models yields much improved designs, typically with very high efficiencies. However, in some experiments exhibiting strong dependence, designs for marginal models may still be inefficient for conditional modelling. Our asymptotic results provide some theoretical insights into why such inefficiencies occur.

Efficient estimation of the number of false positives in high-throughput screening P. 695-704
Holger Rootzén - Dmitrii Zholud

Abstract
This paper develops tail estimation methods to handle false positives in multiple testing problems where testing is done at extreme significance levels and with low degrees of freedom, and where the true null distribution may differ from the theoretical one. We show that the number of false positives, conditional on the total number of positives, has an approximately binomial distribution, and we find estimators of the distribution parameter. We also develop methods for estimation of the true null distribution, as well as techniques to compare it with the theoretical one. Analysis is based on a simple polynomial model for very small *p*-values. Asymptotics that motivate the model, properties of the estimators, and model-checking tools are provided. The methods are applied to two large genomic studies and an fMRI brain scan experiment.

On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions P. 705-711
Jennifer L. Wadsworth

Abstract

Full likelihood-based inference for high-dimensional multivariate extreme value distributions, or max-stable processes, is feasible when incorporating occurrence times of the maxima; without this information, *d*-dimensional likelihood inference is usually precluded due to the large number of terms in the likelihood. However, some studies have noted bias when performing high-dimensional inference that incorporates such event information, particularly when dependence is weak. We elucidate this phenomenon, showing that for unbiased inference in moderate dimensions, dimension *d* should be of a magnitude smaller than the square root of the number of vectors over which one takes the componentwise maximum. A bias reduction technique is suggested and illustrated on the extreme-value logistic model.

Big data and precision P. 712-716
D.R. Cox

Abstract
So-called big data are likely to have complex structure, in particular implying that estimates of precision obtained by applying standard statistical procedures are likely to be misleading, even if the point estimates of parameters themselves may be reasonably satisfactory. While this possibility is best explored in the context of each special case, here we outline a fairly general representation of the accretion of error in large systems and explore the possible implications for the estimation of regression coefficients. The discussion raises issues broadly parallel to the distinction between short-range and long-range dependence in time series theory.

Hysteretic autoregressive time series models P. 717-723
Guodong Li - Bo Guan - Wai Keung Li - Philip L. H. Yu

Abstract
This paper extends the classical two-regime threshold autoregressive model by introducing hysteresis to its regime-switching structure, which leads to a new model: the hysteretic autoregressive model. The proposed model enjoys the piecewise linear structure of a threshold model but has a more flexible regime switching mechanism. A sufficient condition is given for geometric ergodicity. Conditional least squares estimation is discussed, and the asymptotic distributions of its estimators and information criteria for model selection are derived. Simulation results and an example support the model.

Order selection in finite mixture models: complete or observed likelihood information criteria? P. 724-730

Francis K.C. Hui - David I. Warton - and Scott D. Foster

Abstract

Choosing the number of components in a finite mixture model is a challenging task. In this article, we study the behaviour of information criteria for selecting the mixture order, based on either the observed likelihood or the complete likelihood including component labels. We propose a new observed likelihood criterion called aic_{mix}, which is shown to be order consistent. We further show that when there is a nontrivial level of classification uncertainty in the true model, complete likelihood criteria asymptotically underestimate the true number of components. A simulation study illustrates the potentially poor finite-sample performance of complete likelihood criteria, while aic_{mix} and the Bayesian information criterion perform strongly regardless of the level of classification uncertainty.

Sieve maximum likelihood regression analysis of dependent current status data P. 731-738

Ling Ma - Tao Hu - Jianguo Sun

Abstract

Current status data occur in contexts including demographic studies and tumorigenicity experiments. In such cases, each subject is observed only once and the failure time of interest is either left- or right-censored (Kalbfleisch & Prentice, 2002). Many methods have been developed for the analysis of such data (Huang, 1996; Sun, 2006), most of which assume that the failure time and the observation time are independent completely or given covariates. In this paper, we present a sieve maximum likelihood approach for current status data when independence does not hold. A copula model and monotone I-splines are used and the asymptotic properties of the resulting estimators are established. In particular, the estimated regression parameters are shown to be semiparametrically efficient. An illustrative example is provided.

Semiparametric causal inference in matched cohort studies P. 739-746

E. H. Kennedy - A. Sjölander - and D. S. Small

Abstract

Odds ratios can be estimated in case-control studies using standard logistic regression, ignoring the outcome-dependent sampling. In this paper we discuss an analogous result for treatment effects on the treated in matched cohort studies. Specifically, in studies where a sample of treated subjects is observed along with a separate sample of possibly matched controls, we show that efficient and doubly robust estimators of effects on the treated are computationally equivalent to standard estimators, which ignore the matching and exposure-based sampling. This is not the case for general average effects. We also show that matched cohort studies are often more efficient than random sampling for estimating effects on the treated, and derive the optimal number of matches for a given set of matching variables. We illustrate our results via simulation and in a matched cohort study of the effect of hysterectomy on the risk of cardiovascular disease.

A note on convergence of an iterative algorithm for semiparametric odds ratio models P. 747-751

Hua Yun Chen

Abstract

This paper points out an error in Davidov and Iliopoulos's (*Biometrika* 100, 778–80) proof of convergence of an iterative algorithm for the proportional likelihood ratio model. It is shown that the iterative algorithm increases the likelihood in each iteration and converges under mild additional conditions when the odds ratio function is bounded.