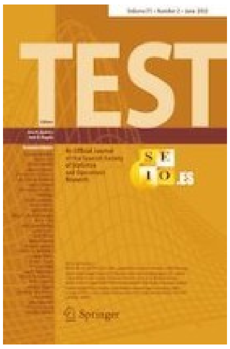

TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 31, number 3 (september 2022)

Penalized robust estimators in sparse logistic regression P. 563-594
Ana M. Bianco, Graciela Boente, Gonzalo Chebi

Abstract
Sparse covariates are frequent in classification and regression problems where the task of variable selection is usually of interest. As it is well known, sparse statistical models correspond to situations where there are only a small number of nonzero parameters, and for that reason, they are much easier to interpret than dense ones. In this paper, we focus on the logistic regression model and our aim is to address robust and penalized estimation for the regression parameter. We introduce a family of penalized weighted *M*-type estimators for the logistic regression parameter that are stable against atypical data. We explore different penalization functions including the so-called Sign penalty. We provide a careful analysis of the estimators convergence rates as well as their variable selection capability and asymptotic distribution for fixed and random penalties. A robust cross-validation criterion is also proposed. Through a numerical study, we compare the finite sample performance of the classical and robust penalized estimators, under different contamination scenarios. The analysis of real datasets enables to investigate the stability of the penalized estimators in the presence of outliers.

Tractable circula densities from Fourier series P. 595-618
Shogo Kato, Arthur Pewsey, M. C. Jones

Abstract
This article proposes an approach, based on infinite Fourier series, to constructing tractable densities for the bivariate circular analogues of copulas recently coined 'circulas'. As examples of the general approach, we consider circula densities generated by various patterns of nonzero Fourier coefficients. The shape and sparsity of such arrangements are found to play a key role in determining the properties of the resultant models. The special cases of the circula densities we consider all have simple closed-form expressions involving no computationally demanding normalizing constants and display wide-ranging distributional shapes. A highly successful model identification tool and methods for parameter estimation and goodness-of-fit testing are provided for the circula densities themselves and the bivariate circular densities obtained from them using a marginal specification construction. The modelling capabilities of such bivariate circular densities are compared with those of five existing models in a numerical experiment, and their application illustrated in an analysis of wind directions.

Weight smoothing for nonprobability surveys P. 619-643
Ramón Ferri-García, Jean-François Beaumont, Kenneth Chu

Abstract
Adjustment techniques to mitigate selection bias in nonprobability samples often involve modelling the propensity to participate in the nonprobability sample along with inverse propensity weighting. It is well known that procedures for estimating weights are effective if the covariates selected in the propensity model are related to both the variable of interest and the participation indicator. In most surveys, there are many variables of interest, making weight

adjustments difficult to determine as a suitable weight for one variable may be unsuitable for other variables. The standard compromise is to include a large number of covariates in the propensity model but this may increase the variability of the estimates, especially when some covariates are weakly related to the variables of interest. Weight smoothing, developed for probability surveys, could be helpful in these situations. It aims to remove the variability caused by overfit propensity models by replacing the inverse propensity weights with predicted weights obtained using a smoothing model. In this article, we study weight smoothing in the nonprobability survey context, both theoretically and empirically, to understand its effectiveness at improving the efficiency of estimates.

A class of random fields with two-piece marginal distributions for modeling point-referenced data with spatial outliers P. 644-674

Moreno Bevilacqua, Christian Caamaño-Carrillo, Camilo Gómez

Abstract

In this paper, we propose a new class of non-Gaussian random fields named two-piece random fields. The proposed class allows to generate random fields that have flexible marginal distributions, possibly skewed and/or heavy-tailed and, as a consequence, has a wide range of applications. We study the second-order properties of this class and provide analytical expressions for the bivariate distribution and the associated correlation functions. We exemplify our general construction by studying two examples: two-piece Gaussian and two-piece Tukey-*h* random fields. An interesting feature of the proposed class is that it offers a specific type of dependence that can be useful when modeling data displaying spatial outliers, a property that has been somewhat ignored from modeling viewpoint in the literature for spatial point referenced data. Since the likelihood function involves analytically intractable integrals, we adopt the weighted pairwise likelihood as a method of estimation. The effectiveness of our methodology is illustrated with simulation experiments as well as with the analysis of a georeferenced dataset of mean temperatures in Middle East.

Data-driven portmanteau tests for time series P. 675-698

Roberto Baragona, Francesco Battaglia, Domenico Cucina

Abstract

Portmanteau tests and information criteria are widely used for checking the hypothesis of independence in time series. More recently, data-driven versions were proposed, where the tests are calibrated based on the largest estimated autocorrelation. It seems natural to introduce a double test statistic (M, Q) where Q is the portmanteau and M is the largest squared autocorrelation. Both statistics have been investigated at length in the past decades. We computed under reasonable assumptions the bivariate probability distribution of this double statistic, conditional, in addition, to the lag at which the largest autocorrelation is found. Tests of the null hypothesis of independence based on rejection regions in the plane (M, Q) are proposed, and some methods to select the rejection region in order to maximize power when the alternative hypothesis is unknown are suggested. A simulation study and a thorough comparison with some popular tests have been performed to show the advantages of our proposal. Notice that this latter includes some well-known univariate tests, so we could expect not only an optimal choice but also additional information which may turn useful for a better understanding of the time series for both model building and forecasting.

General dependence structures for some models based on exponential families with quadratic variance functions P. 699-716

Luis Nieto-Barajas, Eduardo Gutiérrez-Peña

Abstract

We describe a procedure to introduce general dependence structures on a set of random variables. These include order-*q* moving average-type structures, as well as seasonal, periodic, spatial and spatio-temporal dependences. The invariant marginal distribution can be in any family that is conjugate to an exponential family with quadratic variance function. Dependence is induced via a set of suitable latent variables whose conditional distribution mirrors the sampling distribution in a Bayesian conjugate analysis of such exponential families. We obtain strict stationarity as a special case.

On automatic kernel density estimate-based tests for goodness-of-fit

P. 717-748

Carlos Tenreiro

Abstract

Although estimation and testing are different statistical problems, if we want to use a test statistic based on the Parzen–Rosenblatt estimator to test the hypothesis that the underlying density function f is a member of a location-scale family of probability density functions, it may be found reasonable to choose the smoothing parameter in such a way that the kernel density estimator is an effective estimator of f irrespective of which of the null or the alternative hypothesis is true. In this paper we address this question by considering the well-known Bickel–Rosenblatt test statistics which are based on the quadratic distance between the nonparametric kernel estimator and two parametric estimators of f under the null hypothesis. For each one of these test statistics we describe their asymptotic behaviours for a general data-dependent smoothing parameter, and we state their limiting Gaussian null distribution and the consistency of the associated goodness-of-fit test procedures for location-scale families. In order to compare the finite sample power performance of the Bickel–Rosenblatt tests based on a null hypothesis-based bandwidth selector with other bandwidth selector methods existing in the literature, a simulation study for the normal, logistic and Gumbel null location-scale models is included in this work.

Testing marginal homogeneity in Hilbert spaces with applications to stock market returns

P. 749-770

Marc Ditzhaus, Daniel Gaigall

Abstract

This paper considers a paired data framework and discusses the question of marginal homogeneity of bivariate high-dimensional or functional data. The related testing problem can be endowed into a more general setting for paired random variables taking values in a general Hilbert space. To address this problem, a Cramér–von-Mises type test statistic is applied and a bootstrap procedure is suggested to obtain critical values and finally a consistent test. The desired properties of a bootstrap test can be derived that are asymptotic exactness under the null hypothesis and consistency under alternatives. Simulations show the quality of the test in the finite sample case. A possible application is the comparison of two possibly dependent stock market returns based on functional data. The approach is demonstrated based on historical data for different stock market indices.

Increasing the replicability for linear models via adaptive significance levels

P. 771-789

D. Vélez, M. E. Pérez, L. R. Pericchi

Abstract

We put forward an adaptive α (type I error) that decreases as the information grows for hypothesis tests comparing nested linear models. A less elaborate adaptation was presented in Pérez and Pericchi (Stat Probab Lett 85:20–24, 2014) for general i.i.d. models. The calibration proposed in this paper may be interpreted as a Bayes–non-Bayes compromise, of a simple translation of a Bayes factor on frequentist terms that leads to statistical consistency, and most importantly, it is a step toward statistics that promotes replicable scientific findings.

A simple and useful regression model for fitting count data

P. 790-827

Marcelo Bourguignon, Rodrigo M. R. de Medeiros

Abstract

We present a novel regression model for count data where the response variable is BerG-distributed using a new parameterization of this distribution, which is indexed by mean and dispersion parameters. An attractive feature of this model lies in its potential to fit count data when overdispersion, equidispersion, underdispersion, or zero inflation (or deflation) is indicated. The advantage of our new parameterization and approach is the straightforward interpretation of the regression coefficients in terms of the mean and dispersion as in generalized linear models. The maximum likelihood method is used to estimate the model parameters. Also, we conduct hypothesis tests for the dispersion parameter and consider residual analysis. Simulation studies are conducted to empirically evidence the properties of the estimators, the test statistics, and the residuals in finite-sized samples. The proposed model is applied to two real datasets on wildlife habitat and road traffic accidents, which illustrates its capabilities in accommodating both over- and underdispersed count data. This paper contains Supplementary Material.

On finite mixtures of Discretized Beta model for ordered responses P. 828-855

Rosaria Simone

Abstract

The paper discusses the specification of finite mixture models based on the Discretized Beta distribution for the analysis of ordered discrete responses, as ratings and count data. The ultimate goal of the paper is to parameterize clusters of opposite and intermediate response outcomes. After a thorough discussion on model interpretation, identifiability and estimation, the proposal is illustrated on the wake of a case study on the probability to vote for German Political Parties and with a comparative discussion with the state of the art.

Some results on the Gaussian Markov Random Field construction problem based on the use of invariant subgraphs P. 856-874

Juan Baz, Irene Díaz, Raúl Pérez-Fernández

Abstract

The study of Gaussian Markov Random Fields has attracted the attention of a large number of scientific areas due to its increasing usage in several fields of application. Here, we consider the construction of Gaussian Markov Random Fields from a graph and a positive-definite matrix, which is closely related to the problem of finding the Maximum Likelihood Estimator of the covariance matrix of the underlying distribution. In particular, it is simultaneously required that the variances and the covariances between variables associated with adjacent nodes in the graph are fixed by the positive-definite matrix and that pairs of variables associated with non-adjacent nodes in the graph are conditionally independent given all other variables. The solution to this construction problem exists and is unique up to the choice of a vector of means. In this paper, some results focusing on a certain type of subgraphs (invariant subgraphs) and a representation of the Gaussian Markov Random Field as a Multivariate Gaussian Markov Random Field are presented. These results ease the computation of the solution to the aforementioned construction problem.