Biblioteca del Instituto de Estadística y Cartografía de Andalucía

Resúmenes de revistas

Octubre - diciembre 2022

Junta de Andalucía

Consejería de Transformación Económica, Industria, Conocimiento y Universidades

INSTITUTO DE ESTADÍSTICA Y CARTOGRAFÍA DE ANDALUCÍA

PRESENTACIÓN

El presente boletín de resúmenes tiene una periodicidad trimestral y con él la Biblioteca del Instituto de Estadística y Cartografía de Andalucía pretende dar a conocer a los usuarios de una forma detallada el contenido de las revistas especializadas que entran en su colección. Se trata de un complemento al boletín de novedades de publicaciones seriadas ya que en él se incluyen los resúmenes de cada uno de los artículos que aparecen publicados en los diferentes números de las revistas en el idioma original de las mismas.

Los resúmenes de este boletín corresponden a las revistas que han ingresado en la Biblioteca del Instituto de Estadística y Cartografía de Andalucía durante el período **de octubre a diciembre de 2022** y que pueden consultarse gratuitamente en sus instalaciones en la siguiente dirección:

Instituto de Estadística y Cartografía de Andalucía
Pabellón de Nueva Zelanda
C/Leonardo Da Vinci, n. 21. Isla de La Cartuja
41071 - SEVILLA
E-mail: biblio.ieca@juntadeandalucia.es
Teléfono: 955 033 800
Fax: 955 033 816

Horario de atención al público:
Lunes y jueves: de 9:00h a 14:00h. y de 16:00 a 19:00 h.
Martes, miércoles y viernes: de 9:00h a 14:00h.
Horario de verano (del 15 de junio al 15 de septiembre), Semana Santa, Feria de Sevilla y
Navidad (del 24 de diciembre al 6 de enero): de lunes a viernes de 9:00h. a 14:00h.

Technometrics, ISSN 0040-1706
Volume 64, number 1 (2022)

Abstract

Recent developments of advanced imaging systems spur their applications in many areas, ranging from satellite remote sensing for geographic information to thermal imaging analysis for manufacturing process monitoring and control. Due to different specifications of imaging systems, the resulting image stream data (videos) have different spatial and temporal resolutions. This proposed work is based on the image stream data captured by multiple imaging systems for the same object with different but complementary spatial and temporal resolutions. For example, one system has high spatial but low temporal resolutions while the other one has opposite resolutions. The goal of this article is to develop a new super resolution method that integrates these different types of image stream data to improve both spatial and temporal resolutions, which is critical to obtaining more insightful information for more effective quality control of targeted processes or systems. To fulfill this goal, a new tensor completion model is developed by considering both smooth and sparse features simultaneously and is thus termed smooth and sparse tensor completion (SSTC). The results of the extensive case studies illustrate the superiority of our method over the elaborately selected benchmark methods.

Abstract

We consider a binary classification problem in the case where some observations in the training data are incorrectly labeled. In the presence of such label noise, conventional classification fails to obtain a classifier to be generalized to a population. In this work, we investigate label noise logistic regression and explain how it works with noisy training data. We demonstrate that, when label transition probabilities are correctly provided, label noise logistic regression satisfies the Fisher consistency and enjoys the property of robustness. To accommodate various label noise mechanisms that occur in practice, we propose a flexible label noise model in a nonparametric way. We propose an efficient algorithm under the thresholding rule for individual parameter estimation. We demonstrate its performance under synthetic and real examples. We discuss the proposed flexible transition model is also useful for robust classification.

Abstract

Modern scientific instruments readily record various dynamical phenomena at high frequency and for extended durations. Spanning timescales across several orders of magnitude, such “high-throughput” (HTP) data are routinely analyzed with parametric models in the frequency domain. However, the large size of HTP datasets can render maximum likelihood estimation prohibitively expensive. Moreover, HTP recording devices are operated by extensive electronic circuitry, producing periodic noise to which parameter estimates are highly sensitive. This article proposes to address these issues with a two-stage approach. Preliminary parameter estimates are first obtained by a periodogram variance-stabilizing procedure, for which data compression greatly reduces computational costs with minimal impact to statistical efficiency. Next, a novel test with false discovery rate control eliminates most periodic outliers, to which the second-stage estimator becomes more robust. Extensive simulations and experimental results indicate that for a widely used model in HTP data analysis, a substantial reduction in mean squared error can be expected by applying our methodology.

High-Dimensional Cost-constrained Regression Via Nonconvex Optimization P. 52-64

Guan Yu, Haoda Fu & Yufeng Liu

Abstract

Budget constraints become an important consideration in modern predictive modeling due to the high cost of collecting certain predictors. This motivates us to develop cost-constrained predictive modeling methods. In this article, we study a new high-dimensional cost-constrained linear regression problem, that is, we aim to find the cost-constrained regression model with the smallest expected prediction error among all models satisfying a budget constraint. The nonconvex budget constraint makes this problem NP-hard. In order to estimate the regression coefficient vector of the cost-constrained regression model, we propose a new discrete first-order continuous optimization method. In particular, our method delivers a series of estimates of the regression coefficient vector by solving a sequence of 0-1 knapsack problems. Theoretically, we prove that the series of the estimates generated by our iterative algorithm converge to a first-order stationary point, which can be a globally optimal solution under some conditions. Furthermore, we study some extensions of our method that can be used for general statistical learning problems and problems with groups of variables. Numerical studies using simulated datasets and a real dataset from a diabetes study indicate that our proposed method can solve problems of fairly high dimensions with promising performance.

Computer Model Emulation with High-Dimensional Functional Output in Large-Scale Observing System Uncertainty Experiments P. 65-79

Pulong Ma, Anirban Mondal, Bledar A. Konomi, Jonathan Hobbs, Joon Jin Song & Emily L. Kang

Abstract

Observing system uncertainty experiments (OSUEs) have been recently proposed as a cost-effective way to perform probabilistic assessment of retrievals for NASA’s Orbiting Carbon Observatory-2 (OCO-2) mission. One important component in the OCO-2 retrieval algorithm is a full-physics forward model that describes the mathematical relationship between atmospheric variables such as carbon dioxide and radiances measured by the remote sensing instrument. This complex forward model is computationally expensive but large-scale OSUEs require evaluation of this model numerous times, which makes it infeasible for comprehensive experiments. To tackle this issue, we develop a statistical emulator to facilitate large-scale OSUEs in the OCO-2 mission. Within each distinct spectral band, the emulator represents radiance output at irregular wavelengths as a linear combination of basis functions and random coefficients. These random coefficients are then modeled with nearest-neighbor Gaussian processes with built-in input dimension reduction via active subspace. The proposed emulator reduces dimensionality in both input space and output space, so that fast computation is achieved within a fully Bayesian inference framework. Validation experiments demonstrate that this emulator outperforms other competing statistical methods and a reduced order model that approximates the full-physics forward model.

An Information Geometry Approach to Robustness Analysis for the Uncertainty Quantification of Computer Codes P. 80-91

Clement Gauchy, Jerome Stenger, Roman Sueur & Bertrand Iooss

Abstract

Robustness analysis is an emerging field in the uncertainty quantification domain. It involves analyzing the response of a computer model—which has inputs whose exact values are unknown—to the perturbation of one or several of its input distributions. Practical robustness analysis methods therefore require a coherent methodology for perturbing distributions; we present here one such rigorous method, based on the Fisher distance on manifolds of probability distributions. Further, we provide a numerical method to calculate perturbed densities in practice which comes from Lagrangian mechanics and involves solving a system of ordinary differential equations. The method introduced for perturbations is then used to compute quantile-related robustness indices. We illustrate these "perturbed-law based" indices on several numerical models. We also apply our methods to an industrial setting: the simulation of a loss of coolant accident in a nuclear reactor, where several dozen of the model's physical parameters are not known exactly, and where limited knowledge on their distributions is available.

Gaussian Process-Aided Function Comparison Using Noisy Scattered Data P. 92-102

Abhinav Prakash, Rui Tuo & Yu Ding

Abstract

This work proposes a nonparametric method to compare the underlying mean functions given two noisy datasets. The motivation for the work stems from an application of comparing wind turbine power curves. Comparing wind turbine data presents new problems, namely the need to identify the regions of difference in the input space and to quantify the extent of difference that is statistically significant. Our proposed method, referred to as funGP, estimates the underlying functions for different data samples using Gaussian process models. We build a confidence band using the probability law of the estimated function differences under the null hypothesis. Then, the confidence band is used for the hypothesis test as well as for identifying the regions of difference. This identification of difference regions is a distinct feature, as existing methods tend to conduct an overall hypothesis test stating whether two functions are different. Understanding the difference regions can lead to further practical insights and help devise better control and maintenance strategies for wind turbines. The merit of funGP is demonstrated by using three simulation studies and four real wind turbine datasets.

Accounting for Location Measurement Error in Imaging Data With Application to Atomic Resolution Images of Crystalline Materials P. 103-113

Matthew J. Miller, Matthew J. Cabral, Elizabeth C. Dickey, James M. LeBeau & Brian J. Reich

Abstract

Scientists use imaging to identify objects of interest and infer properties of these objects. The locations of these objects are often measured with error, which when ignored leads to biased parameter estimates and inflated variance. Current measurement error methods require an estimate or knowledge of the measurement error variance to correct these estimates, which may not be available. Instead, we create a spatial Bayesian hierarchical model that treats the locations as parameters, using the image itself to incorporate positional uncertainty. We lower the computational burden by approximating the likelihood using a noncontiguous block design around the object locations. We use this model to quantify the relationship between the intensity and displacement of hundreds of atom columns in crystal structures directly imaged via scanning transmission electron microscopy (STEM). Atomic displacements are related to important phenomena such as piezoelectricity, a property useful for engineering applications like ultrasound. Quantifying the sign and magnitude of this relationship will help materials scientists more precisely design materials with improved piezoelectricity. A simulation study confirms our method corrects bias in the estimate of the parameter of interest and drastically improves coverage in high noise scenarios compared to non-measurement error models.

Industrial Forecasting with Exponentially Smoothed Recurrent Neural Networks

Matthew Dixon

Abstract

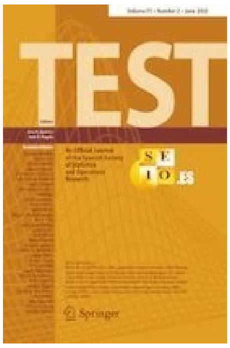
Time series modeling has entered an era of unprecedented growth in the size and complexity of data which require new modeling approaches. While many new general purpose machine learning approaches have emerged, they remain poorly understand and irreconcilable with more traditional statistical modeling approaches. We present a general class of exponential smoothed recurrent neural networks (RNNs) which are well suited to modeling nonstationary dynamical systems arising in industrial applications. In particular, we analyze their capacity to characterize the nonlinear partial autocorrelation structure of time series and directly capture dynamic effects such as seasonality and trends. Application of exponentially smoothed RNNs to forecasting electricity load, weather data, and stock prices highlight the efficacy of exponential smoothing of the hidden state for multistep time series forecasting. The results also suggest that popular, but more complicated neural network architectures originally designed for speech processing are likely over-engineered for industrial forecasting and light-weight exponentially smoothed architectures, trained in a fraction of the time, capture the salient features while being superior and more robust than simple RNNs and autoregressive models. Additionally, uncertainty quantification of Bayesian exponential smoothed RNNs is shown to provide improved coverage.

Prediction of Future Failures for Heterogeneous Reliability Field Data

Colin Lewis-Beck, Qinglong Tian & William Q. Meeker

Abstract

This article introduces methods for constructing prediction bounds or intervals for the number of future failures from heterogeneous reliability field data. We focus on within-sample prediction where early data from a failure-time process is used to predict future failures from the same process. Early data from high-reliability products, however, often have limited information due to some combination of small sample sizes, censoring, and truncation. In such cases, we use a Bayesian hierarchical model to model jointly multiple lifetime distributions arising from different subpopulations of similar products. By borrowing information across subpopulations, our method enables stable estimation and the computation of corresponding prediction intervals, even in cases where there are few observed failures. Three applications are provided to illustrate this methodology, and a simulation study is used to validate the coverage performance of the prediction intervals.



TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 31, number 2 (june 2022)

Multidimensional specification test based on non-stationary time series P. 348-372
Jun Wang, Dianpeng Wang, Yubin Tian

Abstract
In the literature, most works of the specification tests focus on the problem with one-dimensional response or fixed multidimensional responses. In this paper, we develop a new specification test for the parametric models with non-stationary regressor under multidimensional setup, where the dimension of responses may tend to infinity, which fills a gap in the literature. The theoretical results about the asymptotic properties of the proposed test are studied and the optimal rate of the local departure under the alternative hypothesis is also given which ensures the models underpinning by the null and alternative hypotheses can be differentiated. Some simulation studies are done to evaluate the performance of the proposed test with the finite sample. Besides, a real data example based on the US aggregate consumers' consumption data is employed to illustrate the performance. The results of simulation studies and real data analysis both demonstrate the efficiency of our proposed method.

Asymptotics for M-type smoothing splines with non-smooth objective functions P. 373-389
Ioannis Kalogridis

Abstract
M-type smoothing splines are a broad class of spline estimators that include the popular least-squares smoothing spline but also spline estimators that are less susceptible to outlying observations and model misspecification. However, available asymptotic theory only covers smoothing spline estimators based on smooth objective functions and consequently leaves out frequently used resistant estimators such as quantile and Huber-type smoothing splines. We provide a general treatment in this paper and, assuming only the convexity of the objective function, show that the least-squares (super-)convergence rates can be extended to M-type estimators whose asymptotic properties have not been hitherto described. We further show that auxiliary scale estimates may be handled under significantly weaker assumptions than those found in the literature and we establish optimal rates of convergence for the derivatives, which have not been obtained outside the least-squares framework. A simulation study and a real-data example illustrate the competitive performance of non-smooth M-type splines in relation to the least-squares spline on regular data and their superior performance on data that contain anomalies.

Testing the equality of multivariate means when $p>n$ by combining the Hotelling and Simes tests P. 390-415
Tzviel Frostig, Yoav Benjamini

Abstract
We propose a method of testing a shift between mean vectors of two multivariate Gaussian random variables in a high-dimensional setting incorporating the possible dependency and allowing $p>n$. This method is a combination of two well-known tests: the Hotelling test and the Simes test. The tests are integrated by sampling several dimensions at each iteration, testing each using the Hotelling test, and combining their results using the Simes test. We prove that

this procedure is valid asymptotically. This procedure can be extended to handle non-equal covariance matrices by plugging in the appropriate extension of the Hotelling test. Using a simulation study, we show that the proposed test is advantageous over state-of-the-art tests in many scenarios and robust to violation of the Gaussian assumption.

Abstract

Ecological Risk Assessment faces the challenge of determining the impact of invasive species on biodiversity conservation. Although many statistical methods have emerged in recent years in order to model the evolution of the spatio-temporal distribution of invasive species, the notion of extent of occurrence, formally defined by the International Union for the Conservation of Nature, has not been properly handled. In this work, a novel and flexible reconstruction of the extent of occurrence from occurrence data will be established from nonparametric support estimation theory. Mathematically, given a random sample of points from some unknown distribution, we establish a new data-driven method for estimating its probability support S in general dimension. Under the mild geometric assumption that S is r–convex, the smallest r–convex set which contains the sample points is the natural estimator. A stochastic algorithm is proposed for determining an optimal estimate of r from the data under regularity conditions on the density function. The performance of this estimator is studied by reconstructing the extent of occurrence of an assemblage of invasive plant species in the Azores archipelago.

Abstract

Composite quantile regression (CQR) is a robust and efficient estimation method. This paper studies CQR method for single-index models with ultra-high-dimensional data. We propose a penalized CQR estimator for single-index models and combine the debiasing technique with the CQR method to construct an estimator that is asymptotically normal, which enables the construction of valid confidence intervals and hypothesis testing. Both simulations and data analysis are conducted to illustrate the finite sample performance of the proposed methods.

Abstract

Recently, the characterization-based approach for the construction of goodness of fit tests has become popular. Most of the proposed tests have been designed for complete i.i.d. samples. Here, we present the adaptation of the recently proposed exponentiality tests based on equidistribution-type characterizations for the case of randomly censored data. Their asymptotic properties are provided. Besides, we present the results of wide empirical power study including the powers of several recent competitors. This study can be used as a benchmark for future tests proposed for this kind of data.

Abstract

We propose tests for the null hypothesis that the law of a complex-valued random vector is circularly symmetric. The test criteria are formulated as $L2$-type criteria based on empirical characteristic functions, and they are convenient from the computational point of view. Asymptotic as well as Monte Carlo results are presented. Applications on real data are also reported. An R package called CircSymTest is available from the authors.

A test for heteroscedasticity in functional linear models P. 519-542

James Cameron, Pramita Bagchi

Abstract

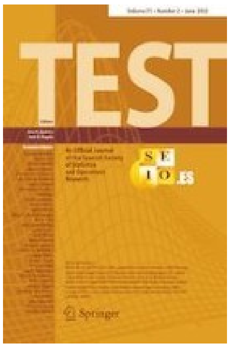
We propose a new test to validate the assumption of homoscedasticity in a functional linear model. We consider a minimum distance measure of heteroscedasticity in functional data, which is zero in the case where the variance is constant and positive otherwise. We derive an explicit form of the measure, propose an estimator for the quantity, and show that an appropriately standardized version of the estimator is asymptotically normally distributed under both the null (homoscedasticity) and alternative hypotheses. We extend this result for residuals from functional linear models and develop a bootstrap diagnostic test for the presence of heteroscedasticity under the postulated model. Moreover, our approach also allows testing for "relevant" deviations from the homoscedastic variance structure and constructing confidence intervals for the proposed measure. We investigate the performance of our method using extensive numerical simulations and a data example.

Stochastic monotonicity of dependent variables given their sum P. 543-561

Franco Pellerey, Jorge Navarro

Abstract

Given a finite set of independent random variables, assume one can observe their sum, and denote with s its value. Efron in 1965, and Lehmann in 1966, described conditions on the involved variables such that each of them stochastically increases in the value s, i.e., such that the expected value of any non-decreasing function of the variable increases as s increases. In this paper, we investigate conditions such that this stochastic monotonicity property is satisfied when the assumption of independence is removed. Comparisons in the stronger likelihood ratio order are considered as well.



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

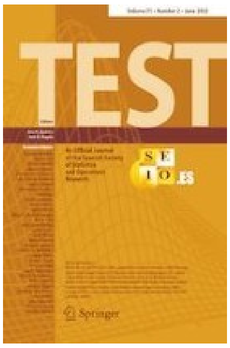
The paper discusses the specification of finite mixture models based on the Discretized Beta distribution for the analysis of ordered discrete responses, as ratings and count data. The ultimate goal of the paper is to parameterize clusters of opposite and intermediate response outcomes. After a thorough 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.



TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 31, number 4 (december 2022)

On sums of dependent random lifetimes under the time-transformed exponential model

P. 879-900

Jorge Navarro, Franco Pellerey, Julio Mulero

Abstract

For a given pair of random lifetimes whose dependence is described by a time-transformed exponential model, we provide analytical expressions for the distribution of their sum. These expressions are obtained by using a representation of the joint distribution in terms of bivariate distortions, which is an alternative approach to the classical copula representation. Since this approach allows one to obtain conditional distributions and their inverses in simple form, then it is also shown how it can be used to predict the value of the sum from the value of one of the variables (or vice versa) by using quantile regression techniques.

Preservation of distributional properties of component lifetimes by system lifetimes

P. 901-930

Barry C. Arnold, Tomasz Rychlik, Magdalena Szymkowiak

Abstract

We analyze reliability systems with components whose lifetimes are identically distributed, and whose joint distribution admits a Samaniego signature representation of the system lifetime distribution. Our main result is the following. We assume that two systems have the same structure and that the lifetimes of the components of the systems share the same dependence copula. If the first system lifetime precedes (succeeds) its single component lifetime in the convex transform order, and if also the component lifetime of the second system precedes the (succeeds) component lifetime of the first system in the convex transform order then the system-component ordering property is preserved by the second system lifetime, i.e., the system lifetime precedes (succeeds) the component lifetime in the second system also. This allows us to conclude various sufficient and necessary conditions on the system signatures under which the monotone failure rate and density properties of the component lifetimes are inherited by the system lifetime under the condition that the component lifetimes are independent.

Testing conditional multivariate rank correlations: the effect of institutional quality on factors influencing competitiveness

P. 931-949

Jone Ascorbebeitia, Eva Ferreira, Susan Orbe

Abstract

Joint distribution between two or more variables could be influenced by the outcome of a conditioning variable. In this paper, we propose a flexible Wald-type statistic to test for such influence. The test is based on a conditioned multivariate Kendall's tau nonparametric estimator. The asymptotic properties of the test statistic are established under different null hypotheses to be tested for, such as conditional independence or testing for constant conditional dependence. Two simulation studies are presented: The first shows that the estimator proposed and the bandwidth selection procedure perform well. The second presents different bivariate and multivariate models to check the size and power of the test and runs comparisons with previous proposals when appropriate. The results support the contention that the test is accurate even in complex situations and that its computational cost is low. As an empirical application, we study the dependence between some pillars of European Regional Competitiveness when conditioned on the quality of regional institutions. We find interesting results, such as weaker links between innovation and higher education in regions with lower institutional quality.

Precision matrix estimation using penalized Generalized Sylvester matrix equation P. 950-967

Vahe Avagyan

Abstract

Estimating a precision matrix is an important problem in several research fields when dealing with large-scale data. Under high-dimensional settings, one of the most popular approaches is optimizing a Lasso or $\ell 1$ norm penalized objective loss function. This penalization endorses sparsity in the estimated matrix and improves the accuracy under a proper calibration of the penalty parameter. In this paper, we demonstrate that the problem of minimizing Lasso penalized D-trace loss can be seen as solving a penalized Sylvester matrix equation. Motivated by this method, we propose estimating the precision matrix using penalized generalized Sylvester matrix equations. In our method, we develop a particular estimating equation and a new convex loss function constructed through this equation, which we call the generalized D-trace loss. We assess the performance of the proposed method using detailed numerical analysis, including simulated and real data. Extensive results show the advantage of the proposed method compared to other estimation approaches in the literature.

Adaptive bi-level variable selection for multivariate failure time model with a diverging number of covariates P. 968-993

Kaida Cai, Hua Shen, Xuewen Lu

Abstract

In this study we propose an adaptive bi-level variable selection method to analyze multivariate failure time data. In the regression setting, we treat the coefficients corresponding to the same predictor variable as a natural group and then consider variable selection at the group level and individual level simultaneously. By imitating the group variable selection procedure with adaptive bi-level penalty, the proposed variable selection method can select a predictor variable at two different levels allowing different covariate effects for different event types: the group level where the predictor is important to all failure types, and the individual level where the predictor is only important to some failure types. An algorithm based on cycle coordinate descent is developed to carry out the proposed method. Based on the simulation results, our method outperforms the classical penalty methods, especially in removing unimportant variables for different failure types. We obtain the asymptotic oracle properties of the proposed variable selection method in the case of a diverging number of covariates. We construct a generalized cross-validation method for the tuning parameter selection and assess model performance using model errors. We also illustrate the proposed method using a real-life data set.

On the general δ-shock model P. 994-1029

Dheeraj Goyal, Nil Kamal Hazra, Maxim Finkelstein

Abstract

he δ-shock model is one of the basic shock models which has a wide range of applications in reliability, finance and related fields. In existing literature, it is assumed that the recovery time of a system from the damage induced by a shock is constant as well as the shocks magnitude. However, as technical systems gradually deteriorate with time, it takes more time to recover from this damage, whereas the larger magnitude of a shock also results in the same effect. Therefore, in this paper, we introduce a general δ-shock model when the recovery time depends on both the arrival times and the magnitudes of shocks. Moreover, we also consider a more general and flexible shock process, namely, the Poisson generalized gamma process. It includes the homogeneous Poisson process, the non-homogeneous Poisson process, the Pólya process and the generalized Pólya process as the particular cases. For the defined survival model, we derive the relationships for the survival function and the mean lifetime and study some relevant stochastic properties. As an application, an example of the corresponding optimal replacement policy is discussed.

Abstract

In this paper, we study the commonness of nonparametric component functions at different quantile levels in spatial additive autoregressive models. We propose two fused adaptive group LASSO penalties to shrink the difference of functions between neighbouring quantile levels. Using these methods, we can estimate the nonparametric functions and identify the quantile regions where functions are unvarying simultaneously. Therefore, when there exists a quantity-level region where the functions are unvarying, its performance is expected to be better than the conventional spatial quantile additive autoregressive model. Under some regularity conditions, the proposed penalized estimators can reach the optimal rate of convergence in theory and also recognize the true varying/unvarying regions accurately. At the same time, our proposed method shows good numerical results in simulated and real datasets.

Abstract

We develop new models for imperfect repair and the corresponding generalized renewal processes for stochastic description of repairable items that fail when their degradation reaches the specified deterministic or random threshold. The discussion is based on the recently suggested notion of a random virtual age and is applied to monotone processes of degradation with independent increments. Imperfect repair reduces degradation of an item on failure to some intermediate level. However, for the nonhomogeneous processes, the corresponding age reduction, which sets back the 'clock' of the process, is also performed. Some relevant stochastic comparisons are obtained. It is shown that the cycles of the corresponding generalized imperfect renewal process are stochastically decreasing/increasing depending on the monotonicity properties of the failure rate that describes the random failure threshold of an item.

Abstract

Modeling bivariate (or multivariate) count data has received increased interest in recent years. The aim is to model the number of different but correlated counts taking into account covariate information. Bivariate Poisson regression models based on the shock model approach are widely used because of their simple form and interpretation. However, these models do not allow for overdispersion or negative correlation, and thus, other models have been proposed in the literature to avoid these limitations. The present paper proposes copula-based bivariate finite mixture of regression models. These models offer some advantages since they have all the benefits of a finite mixture, allowing for unobserved heterogeneity and clustering effects, while the copula-based derivation can produce more flexible structures, including negative correlations and regressors. In this paper, the new approach is defined, estimation through an EM algorithm is presented, and then different models are applied to a Spanish insurance claim count database.

Abstract
Estimation and testing is studied for functional data with temporally dependent errors, an interesting example of which is the event-related potential (ERP). B-spline estimators are formulated for individual smooth trajectories and their population mean as well. The mean estimator is shown to be oracally efficient in the sense that it is as efficient as the infeasible mean estimator if all trajectories had been fully observed without contamination of errors. The oracle efficiency entails asymptotically correct simultaneous confidence band (SCB) for the mean function, which is useful for making inference on the global shape of the mean. Extensive simulation experiments with various time series errors and functional principal components confirm the theoretical conclusions. For a moderate-sized ERP data set, multiple comparison is made by constructing paired SCBs among four different stimuli, over three components N450, N1, and N2 separately or simultaneously, leading to interesting findings.

Abstract
A perennial objection against Bayes factor point-null hypothesis tests is that the point-null hypothesis is known to be false from the outset. We examine the consequences of approximating the sharp point-null hypothesis by a hazy 'peri-null' hypothesis instantiated as a narrow prior distribution centered on the point of interest. The peri-null Bayes factor then equals the point-null Bayes factor multiplied by a correction term which is itself a Bayes factor. For moderate sample sizes, the correction term is relatively inconsequential; however, for large sample sizes, the correction term becomes influential and causes the peri-null Bayes factor to be inconsistent and approach a limit that depends on the ratio of prior ordinates evaluated at the maximum likelihood estimate. We characterize the asymptotic behavior of the peri-null Bayes factor and briefly discuss suggestions on how to construct peri-null Bayes factor hypothesis tests that are also consistent.

Abstract
Never better said, a correct diagnosis is crucial for patient recovery. In the eradication of poverty, which is the first of the sustainable development goals (SDGs) established by the United Nations, efforts in the form of social aid and programs will be useless if they are not directed where they are most needed. Nowadays, monitoring the progress on the SDGs is even more urgent after the sanitary crisis, which is reversing the global poverty reduction observed since 1990 and, given that social development funds are always limited, managing them correctly requires disaggregated statistical information on poverty of acceptable quality. But reliable estimates on living conditions are scarce due to sample size limitations of most official surveys. Common small area estimation procedures supplement the survey data with auxiliary data sources to produce more reliable disaggregated estimates than those based solely on the survey data. We describe the traditional as well as recent model-based procedures for obtaining reliable disaggregated estimates of poverty and inequality indicators, discussing their properties from a practical point of view, placing emphasis on real applications and describing software implementations. We discuss results from recent simulation experiments that compare some of the unit-level methods in terms of bias and efficiency, under model- and design-based setups. Finally, we provide some concluding remarks.

TOP : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1134-5764
Volume 30, number 2 (july 2022)

An adaptative bacterial foraging optimization algorithm for solving the MRCPSP with discounted cash flows

P. 221-248

Luis F. Machado-Domínguez, Carlos D. Paternina-Arboleda, Agustín Barrios-Sarmiento

Abstract

In this paper, a metaheuristic solution algorithm for solving the multi-mode resource-constrained project scheduling problem (MRCPSP) with discounted cash flows (MRCPSPDC) is proposed. This problem consists of determining a schedule such that the project is completed, maximizing the project's net present value (NPV) while complying with the delivery deadline. The adaptative bacterial foraging optimization (ABFO) algorithm is a variation of the original bacterial foraging optimization (BFO), which is a nature-inspired metaheuristic optimization algorithm. We implement a version of the chemotactic operator based on a double justification of the activities given the cash flow. This metaheuristic has been tested in the PSPLIB and MMLIB benchmark datasets available in the literature with promising results. Our ABFO algorithm shows excellent performance in all tested instances and provides suitable solutions for the MRCPSP maximizing the NPV.

An approach to characterizing ϵ-solution sets of convex programs

P. 249-269

N. V. Tuyen, C.-F. Wen, T. Q. Son

Abstract

In this paper, we propose an approach to characterizing ϵ-solution sets of convex programs with a given $\epsilon>0$. The results are divided into two parts. The first one is devoted to establishing the expressions of ϵ-solution sets of a class of convex infinite programs. The representation is given based on the study of relationships among the following three sets: the set of Lagrange multipliers corresponding to a given ϵ-solution, the set of ϵ-solutions of the dual problem corresponding, and the set of ϵ-Kuhn–Tucker vectors associated with the problem in consideration. The second one is devoted to some special cases: the ϵ-solution sets of convex programs that have set constraints and the almost ϵ-solution sets of convex programs that have finite convex constraints. Several examples are given.

The DTC (difference of tangentially convex functions) programming: optimality conditions

P. 270-295

F. Mashkoorzadeh, N. Movahedian, S. Nobakhtian

Abstract

We focus on optimality conditions for an important class of nonconvex and nonsmooth optimization problems, where the objective and constraint functions are presented as a difference of two tangentially convex functions. The main contribution of this paper is to clarify several kinds of stationary solutions and their relations, and establish local optimality conditions with a nonconvex feasible set. Finally, several examples are given to illustrate the effectiveness of the obtained results.

Abstract
The energy consumption of large-scale data centers or server clusters is expected to grow significantly in the next couple of years contributing to up to 13% of the worldwide energy demand in 2030. As the involved processing units require a disproportional amount of energy when they are idle, underutilized, or overloaded, balancing the supply of and the demand for computing resources is a key issue to obtain energy-efficient server consolidations. Whereas traditional concepts mostly consider deterministic predictions of the future workloads or only aim at finding approximate solutions, in this article, we propose an exact approach to tackle the problem of assigning jobs with (not necessarily independent) stochastic characteristics to a minimal amount of servers subject to further practically relevant constraints. As a main contribution, the problem under consideration is reformulated as a stochastic bin packing problem with conflicts and modeled by an integer linear program. Finally, this new approach is tested on real-world instances obtained from a Google data center.

Abstract
In this paper, we study the allocation of a fixed cost among different cities involved in a line-shape transport system like a tram line or a railway. The central characteristic of the problem is that the intended cost is not depending on the infrastructure length or the use intensity. Estañ et al. (Ann Oper Res 301(1):81–105, https://doi.org/10.1007/s10479-020-03645-1, 2021) originally introduced the problem and axiomatically studied it. Based on the well-known bankruptcy problem and game, we analyze it by applying two other approaches. First, adding a parameter, we take into account the municipalities revenues in the determination of cost shares. That enables one to transform a fixed cost allocation problem (FCAP) into a well-known bankruptcy one. We propose two bankruptcy problems for FCAP and use the proportional, adjusted proportional, constrained equal awards, constrained equal losses, and Talmud rules to solve it. Then, we define two bankruptcy games corresponding to FCAP and use the Shapley value for cost allocation. The characteristic functions have attractive interpretations; one considers the agents' minimum desire to contribute to the cost, and the other does their minimum expectation from the overall profit. We investigate presented solutions if they meet some fairness and stability properties. Finally, we apply the suggested approaches to a practical problem.

Abstract
This paper presents a sequential partition method for non-cooperative games of bankruptcy problems. Based on the ascending order of claims, two consequential games are introduced, called the *divide-and-choose game* and the *divide-and-object game*. We prove that the unique Nash equilibrium outcome of each game is consistent with the allocation of the constrained equal awards rule.

Abstract
We consider a cooperative game defined by an economic lot-sizing problem with heterogeneous costs over a finite time horizon, in which each firm faces demand for a single product in each period and coalitions can pool orders. The model of cooperation works as follows: ordering channels and holding and backlogging technologies are shared among the members of the coalitions. This implies that each firm uses the best ordering channel and holding technology

provided by the participants in the consortium. That is, they produce, hold inventory, pay backlogged demand and make orders at the minimum cost of the coalition members. Thus, firms aim at satisfying their demand over the planing horizon with minimal operation cost. Our contribution is to show that there exist fair allocations of the overall operation cost among the firms so that no group of agents profit from leaving the consortium. Then we propose a parametric family of cost allocations and provide sufficient conditions for this to be a stable family against coalitional defections of firms. Finally, we focus on those periods of the time horizon that are consolidated and we analyze their effect on the stability of cost allocations.

An exact lexicographic approach for the maximally risk-disjoint/minimal cost path pair problem in telecommunication networks P. 405-425

Marta Pascoal, José Craveirinha, João Clímaco

Abstract

The paper addresses the lexicographically maximal risk-disjoint/minimal cost path pair problem that aims at finding a pair of paths between two given nodes, which is the shortest (in terms of cost) among those that have the fewest risks in common. This problem is of particular importance in telecommunication network design, namely concerning resilient routing models where both a primary and a backup path have to be calculated to minimize the risk of failure of a connection between origin and terminal nodes, in case of failure along the primary path and where bandwidth routing costs should also be minimized. An exact combinatorial algorithm is proposed for solving this problem which combines a path ranking method and a path labelling algorithm. Also an integer linear programming (ILP) formulation is shown for comparison purposes. After a theoretical justification of the algorithm foundations, this is described and tested, together with the ILP procedure, for a set of reference networks in telecommunications, considering randomly generated risks, associated with Shared Risk Link Groups (SRLGs) and arc costs. Both methods were capable of solving the problem instances in relatively short times and, in general, the proposed algorithm was clearly faster than the ILP formulation excepting for the networks with the greatest dimension and connectivity.

TOP : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1134-5764
Volume 30, number 3 (october 2022)

Optimal pricing for electricity retailers based on data-driven consumers' price-response P. 430-464

Authors

Román Pérez-Santalla, Miguel Carrión, Carlos Ruiz

Abstract

In the present work, we tackle the problem of finding the optimal price tariff to be set by a risk-averse electric retailer participating in the pool and whose customers are price sensitive. We assume that the retailer has access to a sufficiently large smart-meter dataset from which it can statistically characterize the relationship between the tariff price and the demand load of its clients. Three different models are analyzed to predict the aggregated load as a function of the electricity prices and other parameters, as humidity or temperature. More specifically, we train linear regression (predictive) models to forecast the resulting demand load as a function of the retail price. Then, we will insert this model in a quadratic optimization problem which evaluates the optimal price to be offered. This optimization problem accounts for different sources of uncertainty including consumer's response, pool prices and renewable source availability, and relies on a stochastic and risk-averse formulation. In particular, one important contribution of this work is to base the scenario generation and reduction procedure on the statistical properties of the resulting predictive model. This allows us to properly quantify (data-driven) not only the expected value but the level of uncertainty associated with the main problem parameters. Moreover, we consider both standard forward-based contracts and the recently introduced power purchase agreement contracts as risk-hedging tools for the retailer. The results are promising as profits are found for the retailer with highly competitive prices and some possible improvements are shown if richer datasets could be available in the future. A realistic case study and multiple sensitivity analyses have been performed to characterize the risk-aversion behavior of the retailer considering price-sensitive consumers. It has been assumed that the energy procurement of the retailer can be satisfied from the pool and different types of contracts. The obtained results reveal that the risk-aversion degree of the retailer strongly influences contracting decisions, whereas the price sensitiveness of consumers has a higher impact on the selling price offered.

Solving certain complementarity problems in power markets via convex programming P. 465-491

G. Constante-Flores, A. J. Conejo, S. Constante-Flores

Abstract

We address the solution of certain Mathematical Programs with Equilibrium Constraints (MPECs) in power markets using convex optimization. These MPECs constitute a class of complementarity problems relevant to the design and operation of power markets. Specifically, given a non-convex continuous MPEC of the considered type, we iteratively solve a collection of convex optimization problems that approximate the MPEC until a pre-specified tolerance is reached. We use an insightful example to illustrate the proposed solution technique and a case study to analyze its computational performance.

Ambiguities and nonmonotonicities under prosumer power P. 492-532

Afzal S. Siddiqui, Sauleh A. Siddiqui

Abstract

Prosumers adopt distributed energy resources (DER) to cover part of their own consumption and to sell surplus energy. Although individual prosumers are too dispersed to exert operational market power, they may collectively hold a strategic advantage over conventional generation in selecting DER capacity via aggregators. We devise a bilevel model to examine DER capacity sizing by a collective prosumer as a Stackelberg leader in an electricity industry where conventional generation may exert market power in operations. At the upper level, the prosumer chooses DER capacity in anticipation of lower-level operations by conventional generation and DER output. We demonstrate that exertion of market power in operations by conventional generation and the marginal cost of conventional generation affect DER investment by the prosumer in a nonmonotonic manner. Intuitively, in an industry where conventional generation exerts market power in operations similar to a monopoly (MO), the prosumer invests in more DER capacity than under perfectly competitive operations (PC) to take advantage of a high market-clearing price. However, if the marginal cost of conventional generation is high enough, then this intuitive result is reversed as the prosumer adopts more DER capacity under PC than under MO. This is because the high marginal cost of conventional generation prevents the market-clearing price from decreasing, thereby allowing for higher prosumer revenues. Moreover, competition relieves the chokehold on consumption under MO, which further incentivises the prosumer to expand DER capacity to capture market share. We prove the existence of a critical threshold for the marginal cost of conventional generation that leads to this counterintuitive result. Finally, we propose a countervailing regulatory mechanism that yields welfare-enhancing DER investment even in deregulated electricity industries.

Regression markets and application to energy forecasting P. 533-573

Authors

Pierre Pinson, Liyang Han, Jalal Kazempour

Abstract

Energy forecasting has attracted enormous attention over the last few decades, with novel proposals related to the use of heterogeneous data sources, probabilistic forecasting, online learning, etc. A key aspect that emerged is that learning and forecasting may highly benefit from distributed data, though not only in the geographical sense. That is, various agents collect and own data that may be useful to others. In contrast to recent proposals that look into distributed and privacy-preserving learning (incentive-free), we explore here a framework called regression markets. There, agents aiming to improve their forecasts post a regression task, for which other agents may contribute by sharing their data for their features and get monetarily rewarded for it. The market design is for regression models that are linear in their parameters, and possibly separable, with estimation performed based on either batch or online learning. Both in-sample and out-of-sample aspects are considered, with markets for fitting models in-sample, and then for improving genuine forecasts out-of-sample. Such regression markets rely on recent concepts within interpretability of machine learning approaches and cooperative game theory, with Shapley additive explanations. Besides introducing the market design and proving its desirable properties, application results are shown based on simulation studies (to highlight the salient features of the proposal) and with real-world case studies.

The impact of convexity on expansion planning in low-carbon electricity markets P. 574–593

S. Wogrin, D. Tejada-Arango, A. Botterud

Abstract

Expansion planning models are tools frequently employed to analyze the transition to a carbon-neutral power system. Such models provide estimates for an optimal technology mix and optimal operating decisions, but they are also often used to obtain prices and subsequently calculate profits. This paper analyzes the impact of modeling assumptions on convexity for power system outcomes and, in particular, on investment cost recovery. Through a case study, we find that although there is a long-term equilibrium for producers under convex models, introducing realistic constraints, such as non-convexities/lumpiness of investments, inelastic demand or unit commitment constraints, leads to profitability challenges. We furthermore demonstrate that considering only short-term marginal costs in market-clearing may potentially create a significant missing-money problem caused by a missing-market problem and dual degeneracy in a 100 percent renewable system.

Integrating unimodality into distributionally robust optimal power flow P. 594–617

Bowen Li, Ruiwei Jiang, Johanna L. Mathieu

Abstract

To manage renewable generation and load consumption uncertainty, chance-constrained optimal power flow (OPF) formulations have been proposed. However, conventional solution approaches often rely on accurate estimates of uncertainty distributions, which are rarely available in reality. When the distributions are not known but can be limited to a set of plausible candidates, termed an ambiguity set, distributionally robust (DR) optimization can reduce out-of-sample violation of chance constraints. Nevertheless, a DR model may yield conservative solutions if the ambiguity set is too large. In view that most practical uncertainty distributions for renewable generation are unimodal, in this paper, we integrate unimodality into a moment-based ambiguity set to reduce the conservatism of a DR-OPF model. We review exact reformulations, approximations, and an online algorithm for solving this model. We extend these results to derive a new, offline solution algorithm. Specifically, this algorithm uses a parameter selection approach that searches for an optimal approximation of the DR-OPF model before solving it. This significantly improves the computational efficiency and solution quality. We evaluate the performance of the offline algorithm against existing solution approaches for DR-OPF using modified IEEE 118-bus and 300-bus systems with high penetrations of renewable generation. Results show that including unimodality reduces solution conservatism and cost without degrading reliability significantly.

Recent contributions to the optimal design of pipeline networks in the energy industry using mathematical programming P. 618-648

Diego C. Cafaro, Demian J. Presser, Ignacio E. Grossmann

Abstract

The optimal design of pipeline networks has inspired process systems engineers and operations research practitioners since the earliest times of mathematical programming. The nonlinear equations governing pressure drops, energy consumption and capital investments have motivated nonlinear programming (NLP) approaches and solution techniques, as well as mixed-integer nonlinear programming (MINLP) formulations and decomposition strategies. In this overview paper, we present a systematic description of the mathematical models proposed in recent years for the optimal design of pipeline networks in the energy industry. We provide a general framework to address these problems based on both the topology of the network to build, and the physical properties of the fluids to transport. We illustrate the computational challenges through several examples from industry collaboration projects, published in recent papers from our research group.

Data-driven tuning for chance constrained optimization: analysis and extensions P. 649-682

Ashley M. Hou, Line A. Roald

Abstract

Many optimization problems involve uncertain parameters which, if not appropriately accounted for, can cause solution infeasiblity. In this work, we consider joint chance-constrained optimization problems, which require all constraints to hold with a given probability, and a two-step solution method based on iterative tuning. Previous work established an a priori feasibility guarantee for this tuning approach, which relies on an assumption that must be verified on a case-by-case basis. In this paper, we propose an empirical methodology using statistical hypothesis testing to assess the validity of this assumption, thus providing further insight into the validity of the a priori guarantee. In addition, we provide sample complexity results to assess the requisite amount of data for the tuning method. We find that for large scale optimization problems, the tuning approach may require significantly less samples than the scenario approach.

We numerically assess these results via application to the optimal power flow problem as well as further assess the scalability of the method and the optimality and feasibility of solutions obtained from tuning.

Day-ahead market bidding taking the balancing power market into account p. 683-703

Authors (first, second and last of 4)
Gro Klæboe, Jørgen Braathen, Stein-Erik Fleten

Abstract
Generation companies with controllable units put considerable analysis into the process of bidding into the day-ahead markets for electricity. This article investigates the gain of coordinating price-taking bids to the day-ahead electricity market (DA) and sequentially cleared energy-only markets, such as the Nordic balancing market (BM). A technically detailed case study from the Nordic market is presented. We find that coordinated bidding is hardly worthwhile under current market conditions, but that only a modest increase in the demand for balancing energy will make coordination profitable. If the supply curve for balancing energy is convex, so that the cost of balancing energy is asymmetric, the gains will be even higher. Finally, we find that day-ahead market bid curves that result from coordinated instances provide extra supply at low prices, and lower supply at high prices, compared to sequential bids. This is rational given the anticipated opportunities that the balancing market offers; however, it makes day-ahead bidding appear to exploit market power.