Biblioteca del Instituto de Estadística y Cartografía de Andalucía

Resúmenes de revistas
Septiembre-octubre 2017

JUNTA DE ANDALUCIA

Instituto de Estadística y Cartografía de Andalucía
CONSEJERÍA DE ECONOMÍA, INNOVACIÓN, CIENCIA Y EMPLEO

PRESENTACIÓN

El presente boletín de resúmenes tiene una periodicidad bimestral 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 septiembre y octubre de 2017 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 martes: de 9:00h a 14:00h. y de 16:00 a 19:00 h.
Miércoles, jueves 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.

Cartographic journal, The, ISSN 0008-7041
Volume 53, number 1 (february 2016)

Abstract
It is well known that all longitudes in Ptolemy's *Geography* are cumulatively overestimated, so that his map is excessively stretched out from west to east in comparison with the modern map. In recent years, a hypothesis have been advanced that this stretching can be explained as a result of the change in the value of the Earth's circumference from a larger one proposed by Eratosthenes to a lesser one by Posidonius. This explanation has two necessary presuppositions: (1) that Ptolemy's map is stretched out by a factor of ~1·4 which coincides with the ratio between Eratosthenes' and Posidonius' values, and (2) that Ptolemy's error in longitude grows almost linearly. This article argues that the situation is more complex and nuanced. In fact, the error in longitude on Ptolemy's map (and the stretching factor of the map, accordingly) varies considerably depending on longitude, latitude, and region. In particular, the error grows most slowly in the Eastern Mediterranean, which is probably due to the fact that this region was the centre of the ancient world. Therefore, Ptolemy's error in longitude cannot be explained by one universal cause, but only by a combination of different factors.

Abstract
A sample of 18th, 19th and 20th-Century historical star atlases from the Royal Institute and Observatory of the Spanish Navy and the Linda Hall Library have been selected in order to identify the most frequently supplied scientific information. This work shows how the quality of bibliographic records could be improved, not only by adding more specialized description fields, but also by ensuring that the existing ones are being properly used by cataloguers. A series of new technical parameters is proposed, along with guidelines on how to find them, thus making the task of identifying such parameters easier for cataloguers.

Abstract
A proper characterization of land-use types is critical for constructing generalization constraints to guide and control landuse data generalization. This paper focused on identification and utilization of their importance based upon land-use distributions and application themes. First, this importance was identified using a three-step method that links a diversity index, a multiple attribute decision model and a spatial association analysis. Second, with the importance, a mathematical function was designed to determine minimum area thresholds of land-use polygons as an example of generalization constraints. Third, the importance was used to assist in the selection of generalization operators and evaluation of generalization outcomes. Fourth, a land-use dataset at 1:10 000, describing the land use of a typical rural area in Hubei

province of China, was generalized towards a 1:50 000 dataset to verify the effects of the presented method and function. Three additional tests were implemented to analyze the sensitivity of the importance of land-use types on setting the minimum area threshold and generalization operations. The outcome showed that the proposed methods and functions make land-use data generalization more adaptable for in-use datasets and applications.

The Interaction of Landmarks and Map Alignment in You-Are-Here Maps P. 43-54

Grant McKenzie & Alexander Klippel

Abstract

Knowing where one is located within an environment is one of the most fundamental tasks humans have to master in their daily routines. Maps, as external representations of the environment offer intuitive ways to extend the capacities of the human cognitive systems. Operations such as planning a route can be performed on maps instead of in the environment. Question of how to design maps that support cognitive processes such as wayfinding in novel environments have been discussed in several disciplines. The research reported here addresses the question of how map alignment and the presence of landmarks in maps interact during wayfinding. For the purpose of systematically analyzing the relationship between map alignment and landmark presence, nine virtual environments were designed. Routes learned from maps with different alignments and different numbers of landmarks present at decision points were used. While generally landmarks are assumed to foster wayfinding performance, our results indicate that misaligned maps can cancel out positive effects obtained through landmarks.

Portable Map Signboard Collages on Smartphones Using Polyline Georeferences P. 55-65

Ruochen Si & Masatoshi Arikawa

Abstract

Different types of maps have specific, limited functions, and to satisfy user requirements under various situations, combinations of types are required. With the development and popularization of mapping applications on mobile devices, efforts have been made to integrate analogue maps with location-based services (LBSs). Previous work in this area has mainly used points as georeferences to calculate positions on map images, but point-based geocoding does not achieve accurate and stable positioning results. This paper introduces a map signboard collage system (MSCS) to enable LBSs using copied images of multiple map signboards on smartphones. The system uses polylines as georeferences, improving the accuracy and stability of positioning results, and supports automatic egocentralization of map images with compass sensors in smartphones. In addition, it collages multiple maps to allow natural, smooth and dynamic map switching for inter-map navigation.

Empirical Antecedents of Representation of Relief Features in Plan. The Case of Spanish American Cartography in the Sixteenth Century: Three Significant Examples P. 66-77

Manuel Morato-Moreno

Abstract

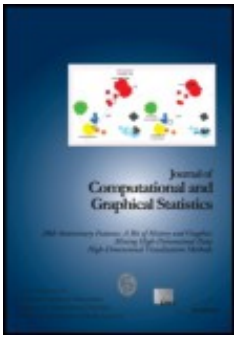
Although the representation of the Earth's surface in plan has been used since antiquity, the first and most important uses of these representations come from the Renaissance. As a result of the discovery of the new American territories, many topographic surveys were performed during the sixteenth century. These maps developed various empirical methods to represent terrain and relief. In this comprehensive corpus, we find three maps with surprising representations of terrain that depart from pictorial methods, such as hill profiles or aligned or grouped mounds (sugar loaf hills), that were used from the origins of cartography until the eighteenth century. The procedure used on these maps is, to a certain extent, an intuitive anticipation of the scientific method used by modern surveying, in which the land relief is represented in an orthogonal view by the contour lines. This method was not developed and systematized until two hundred years after the three maps analysed in this study were produced.

Mapping Space-Related Emotions out of User-Generated Photo Metadata Considering Grammatical Issues P. 78-90

Eva Hauthal & Dirk Burghardt

Abstract

Current location based services mainly provide objective information and collections of facts. Subjective components such as emotions and opinions can provide additional alternative information useful in decision making, e.g. in tourism, business, entertainment and the like. Therefore research on effect analysis was carried out by capturing and analyzing georeferenced emotions from user-generated content. An approach was developed for extracting location-based emotions from the written language in the metadata of georeferenced Flickr and Panoramio photos, i.e. from their titles, descriptions and tags. Within this extraction approach various grammatical issues were considered, like negations of words or amplifications. Procedures were developed for modifying the affected emotions, for example for inverting or intensifying them. The approach was applied to the study area of Dresden, Germany. The obtained emotions were documented in emotional maps of geospace as well as in valence-arousal-space originating from psychology.



Journal of computational and graphical statistics, ISSN 1061-8600
Volume 26, number 2 (2017)

Abstract
Graphics are very effective for communicating numerical information quickly and efficiently, but many of the design choices we make are based on subjective measures, such as personal taste or conventions of the discipline rather than objective criteria. We briefly introduce perceptual principles such as preattentive features and gestalt heuristics, and then discuss the design and results of a factorial experiment examining the effect of plot aesthetics such as color and trend lines on participants' assessment of ambiguous data displays. The quantitative and qualitative experimental results strongly suggest that plot aesthetics have a significant impact on the perception of important features in data displays. Supplementary materials for this article are available online.

Abstract
Graphs are powerful and versatile data structures that can be used to represent a wide range of different types of information. In this article, we introduce a method to analyze and then visualize an important class of data described over a graph—namely, ensembles of paths. Analysis of such path ensembles is useful in a variety of applications, in diverse fields such as transportation, computer networks, and molecular dynamics. The proposed method generalizes the concept of *band depth* to an ensemble of paths on a graph, which provides a center-outward ordering on the paths. This ordering is, in turn, used to construct a generalization of the conventional boxplot or whisker plot, called a *path boxplot*, which applies to paths on a graph. The utility of path boxplot is demonstrated for several examples of path ensembles including paths defined over computer networks and roads. Supplementary materials for this article are available online.

Abstract
We introduce a graphical diagnostic called the Torgegram for characterizing spatial dependence among observations of a variable on a stream network. The Torgegram consists of four component empirical semivariograms, each one corresponding to a particular combination of flow-connectedness within the network and model type (tail-up/tail-down). We show how an overall strategy for fluvial variography can be based on a careful examination of the Torgegram. An analysis of water temperature data from a stream network within the Columbia River basin of the northwest United States illustrates the diagnostic value of the Torgegram as well as its limitations. Additional uses and extensions of the Torgegram are discussed.

Semiparametric Bayesian Regression via Potts Model P. 265-274

Alejandro Murua & Fernando A. Quintana

Abstract

We consider Bayesian nonparametric regression through random partition models. Our approach involves the construction of a covariate-dependent prior distribution on partitions of individuals. Our goal is to use covariate information to improve predictive inference. To do so, we propose a prior on partitions based on the Potts clustering model associated with the observed covariates. This drives by covariate proximity both the formation of clusters, and the prior predictive distribution. The resulting prior model is flexible enough to support many different types of likelihood models. We focus the discussion on nonparametric regression. Implementation details are discussed for the specific case of multivariate multiple linear regression. The proposed model performs well in terms of model fitting and prediction when compared to other alternative nonparametric regression approaches. We illustrate the methodology with an application to the health status of nations at the turn of the 21st century. Supplementary materials are available online.

Regression Adjustment for Noncrossing Bayesian Quantile Regression P. 275-284

T. Rodrigues & Y. Fan

Abstract

A two-stage approach is proposed to overcome the problem in quantile regression, where separately fitted curves for several quantiles may cross. The standard Bayesian quantile regression model is applied in the first stage, followed by a Gaussian process regression adjustment, which monotonizes the quantile function while borrowing strength from nearby quantiles. The two-stage approach is computationally efficient, and more general than existing techniques. The method is shown to be competitive with alternative approaches via its performance in simulated examples. Supplementary materials for the article are available online.

Identifying Mixtures of Mixtures Using Bayesian Estimation P. 285-295

Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter & Bettina Grün

Abstract

The use of a finite mixture of normal distributions in model-based clustering allows us to capture non-Gaussian data clusters. However, identifying the clusters from the normal components is challenging and in general either achieved by imposing constraints on the model or by using post-processing procedures. Within the Bayesian framework, we propose a different approach based on sparse finite mixtures to achieve identifiability. We specify a hierarchical prior, where the hyperparameters are carefully selected such that they are reflective of the cluster structure aimed at. In addition, this prior allows us to estimate the model using standard MCMC sampling methods. In combination with a post-processing approach which resolves the label switching issue and results in an identified model, our approach allows us to simultaneously (1) determine the number of clusters, (2) flexibly approximate the cluster distributions in a semiparametric way using finite mixtures of normals and (3) identify cluster-specific parameters and classify observations. The proposed approach is illustrated in two simulation studies and on benchmark datasets. Supplementary materials for this article are available online.

Combining Functional Data Registration and Factor Analysis P. 296-305

Cecilia Earls & Giles Hooker

Abstract

We extend the definition of functional data registration to encompass a larger class of registration models. In contrast to traditional registration models, we allow for registered functions that have more than one primary direction of variation. The proposed Bayesian hierarchical model simultaneously registers the observed functions and estimates the two primary factors that characterize variation in the registered functions. Each registered function is assumed to be predominantly composed of a linear combination of these two primary factors, and the function-specific weights for each observation are estimated within the registration model. We show how these estimated weights can easily be used to classify functions after registration using both simulated data and a juggling dataset. Supplementary materials

for this article are available online.

Locally Sparse Estimator for Functional Linear Regression Models P. 306-318

Zhenhua Lin, Jiguo Cao, Liangliang Wang & Haonan Wang

Abstract

A new locally sparse (i.e., zero on some subregions) estimator for coefficient functions in functional linear regression models is developed based on a novel functional regularization technique called "fSCAD." The nice shrinkage property of fSCAD allows the proposed estimator to locate null subregions of coefficient functions without over shrinking nonzero values of coefficient functions. Additionally, a roughness penalty is incorporated to control the roughness of the locally sparse estimator. Our method is theoretically sounder and computationally simpler than existing methods. Asymptotic analysis reveals that the proposed estimator is consistent and can identify null subregions with probability tending to one. Extensive simulations confirm the theoretical analysis and show excellent numerical performance of the proposed method. Practical merit of locally sparse modeling is demonstrated by two real applications. Supplemental materials for the article are available online.

Sparse Functional Dynamical Models—A Big Data Approach P. 319-329

Ela Sienkiewicz, Dong Song, F. Jay Breidt & Haonan Wang

Abstract

Nonlinear dynamical systems are encountered in many areas of social science, natural science, and engineering, and are of particular interest for complex biological processes like the spiking activity of neural ensembles in the brain. To describe such spiking activity, we adapt the Volterra series expansion of an analytic function to account for the point-process nature of multiple inputs and a single output (MISO) in a neural ensemble. Our model describes the transformed spiking probability for the output as the sum of kernel-weighted integrals of the inputs. The kernel functions need to be identified and estimated, and both local sparsity (kernel functions may be zero on part of their support) and global sparsity (some kernel functions may be identically zero) are of interest. The kernel functions are approximated by B-splines and a penalized likelihood-based approach is proposed for estimation. Even for moderately complex brain functionality, the identification and estimation of this sparse functional dynamical model poses major computational challenges, which we address with big data techniques that can be implemented on a single, multi-core server. The performance of the proposed method is demonstrated using neural recordings from the hippocampus of a rat during open field tasks. Supplementary materials for this article are available online.

Grouped Functional Time Series Forecasting: An Application to Age-Specific Mortality Rates P. 330-343

Han Lin Shang & Rob J. Hyndman

Abstract

Age-specific mortality rates are often disaggregated by different attributes, such as sex, state, and ethnicity. Forecasting age-specific mortality rates at the national and sub-national levels plays an important role in developing social policy. However, independent forecasts at the sub-national levels may not add up to the forecasts at the national level. To address this issue, we consider reconciling forecasts of age-specific mortality rates, extending the methods of Hyndman et al. in 2011 Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011), "Optimal Combination Forecasts for Hierarchical Time Series," *Computational Statistics and Data Analysis*, 55, 2579–2589. [Crossref], [Web of Science ®], [Google Scholar] to functional time series, where age is considered as a continuum. The grouped functional time series methods are used to produce point forecasts of mortality rates that are aggregated appropriately across different disaggregation factors. For evaluating forecast uncertainty, we propose a bootstrap method for reconciling interval forecasts. Using the regional age-specific mortality rates in Japan, obtained from the Japanese Mortality Database, we investigate the one- to ten-step-ahead point and interval forecast accuracies between the independent and grouped functional time series forecasting methods. The proposed methods are shown to be

useful for reconciling forecasts of age-specific mortality rates at the national and sub-national levels. They also enjoy improved forecast accuracy averaged over different disaggregation factors. Supplementary materials for the article are available online.

A Semiparametric Two-Sample Hypothesis Testing Problem for Random Graphs P. 344-354

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park & Carey E. Priebe

Abstract

Two-sample hypothesis testing for random graphs arises naturally in neuroscience, social networks, and machine learning. In this article, we consider a semiparametric problem of two-sample hypothesis testing for a class of latent position random graphs. We formulate a notion of consistency in this context and propose a valid test for the hypothesis that two finite-dimensional random dot product graphs on a common vertex set have the same generating latent positions or have generating latent positions that are scaled or diagonal transformations of one another. Our test statistic is a function of a spectral decomposition of the adjacency matrix for each graph and our test procedure is consistent across a broad range of alternatives. We apply our test procedure to real biological data: in a test-retest dataset of neural connectome graphs, we are able to distinguish between scans from different subjects; and in the *C. elegans* connectome, we are able to distinguish between chemical and electrical networks. The latter example is a concrete demonstration that our test can have power even for small-sample sizes. We conclude by discussing the relationship between our test procedure and generalized likelihood ratio tests. Supplementary materials for this article are available online.

Sparse Steinian Covariance Estimation P. 355-366

Brett Naul & Jonathan Taylor

Abstract

We consider a new method for sparse covariance matrix estimation which is motivated by previous results for the so-called Stein-type estimators. Stein proposed a method for regularizing the sample covariance matrix by shrinking together the eigenvalues; the amount of shrinkage is chosen to minimize an unbiased estimate of the risk (UBEOR) under the entropy loss function. The resulting estimator has been shown in simulations to yield significant risk reductions over the maximum likelihood estimator. Our method extends the UBEOR minimization problem by adding an ℓ_1 penalty on the entries of the estimated covariance matrix, which encourages a sparse estimate. For a multivariate Gaussian distribution, zeros in the covariance matrix correspond to marginal independences between variables. Unlike the ℓ_1-penalized Gaussian likelihood function, our penalized UBEOR objective is convex and can be minimized via a simple block coordinate descent procedure. We demonstrate via numerical simulations and an analysis of microarray data from breast cancer patients that our proposed method generally outperforms other methods for sparse covariance matrix estimation and can be computed efficiently even in high dimensions.

High-Dimensional Mixed Graphical Models P. 367-378

Jie Cheng, Tianxi Li, Elizaveta Levina & Ji Zhu

Abstract

While graphical models for continuous data (Gaussian graphical models) and discrete data (Ising models) have been extensively studied, there is little work on graphical models for datasets with both continuous and discrete variables (mixed data), which are common in many scientific applications. We propose a novel graphical model for mixed data, which is simple enough to be suitable for high-dimensional data, yet flexible enough to represent all possible graph structures. We develop a computationally efficient regression-based algorithm for fitting the model by focusing on the conditional log-likelihood of each variable given the rest. The parameters have a natural group structure, and sparsity in the fitted graph is attained by incorporating a group lasso penalty, approximated by a weighted lasso penalty for computational efficiency. We demonstrate the effectiveness of our method through an extensive simulation study and apply it to a music annotation dataset (CAL500), obtaining a sparse and interpretable graphical model relating the

continuous features of the audio signal to binary variables such as genre, emotions, and usage associated with particular songs. While we focus on binary discrete variables for the main presentation, we also show that the proposed methodology can be easily extended to general discrete variables.

Penalized Versus Constrained Generalized Eigenvalue Problems P. 379-387

Irina Gaynanova, James G. Booth & Martin T. Wells

Abstract

We investigate the difference between using an ℓ_1 penalty versus an ℓ_1 constraint in generalized eigenvalue problems arising in multivariate analysis. Our main finding is that the ℓ_1 penalty may fail to provide very sparse solutions; a severe disadvantage for variable selection that can be remedied by using an ℓ_1 constraint. Our claims are supported both by empirical evidence and theoretical analysis. Finally, we illustrate the advantages of the ℓ_1 constraint in the context of discriminant analysis and principal component analysis. Supplementary materials for this article are available online.

Composite Likelihood Inference in a Discrete Latent Variable Model for Two-Way "Clustering-by-Segmentation" Problems P. 388-402

Francesco Bartolucci, Francesca Chiaromonte, Prabhani Kuruppumullage Don & Bruce G. Lindsay

Abstract

We consider a discrete latent variable model for two-way data arrays, which allows one to simultaneously produce clusters along one of the data dimensions (e.g., exchangeable observational units or features) and contiguous groups, or segments, along the other (e.g., consecutively ordered times or locations). The model relies on a hidden Markov structure but, given its complexity, cannot be estimated by full maximum likelihood. Therefore, we introduce a composite likelihood methodology based on considering different subsets of the data. The proposed approach is illustrated by simulation, and with an application to genomic data.

Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE P. 403-413

Perry de Valpine, Daniel Turek, Christopher J. Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang & Rastislav Bodik

Abstract

We describe NIMBLE, a system for programming statistical algorithms for general model structures within R. NIMBLE is designed to meet three challenges: flexible model specification, a language for programming algorithms that can use different models, and a balance between high-level programmability and execution efficiency. For model specification, NIMBLE extends the BUGS language and creates model objects, which can manipulate variables, calculate log probability values, generate simulations, and query the relationships among variables. For algorithm programming, NIMBLE provides functions that operate with model objects using two stages of evaluation. The first stage allows specialization of a function to a particular model and/or nodes, such as creating a Metropolis-Hastings sampler for a particular block of nodes. The second stage allows repeated execution of computations using the results of the first stage. To achieve efficient second-stage computation, NIMBLE compiles models and functions via C++, using the Eigen library for linear algebra, and provides the user with an interface to compiled objects. The NIMBLE language represents a compilable domain-specific language (DSL) embedded within R. This article provides an overview of the design and rationale for NIMBLE along with illustrative examples including importance sampling, Markov chain Monte Carlo (MCMC) and Monte Carlo expectation maximization (MCEM). Supplementary materials for this article are available online.

Adaptive Sequential Monte Carlo for Multiple Changepoint Analysis P. 414-423

Nicholas A. Heard & Melissa J. M. Turcotte

Abstract

Process monitoring and control requires the detection of structural changes in a data stream in real time. This article introduces an efficient sequential Monte Carlo algorithm designed for learning unknown changepoints in continuous time. The method is intuitively simple: new changepoints for the latest window of data are proposed by conditioning only on data observed since the most recent estimated changepoint, as these observations carry most of the information about the current state of the process. The proposed method shows improved performance over the current state of the art. Another advantage of the proposed algorithm is that it can be made adaptive, varying the number of particles according to the apparent local complexity of the target changepoint probability distribution. This saves valuable computing time when changes in the changepoint distribution are negligible, and enables rebalancing of the importance weights of existing particles when a significant change in the target distribution is encountered. The plain and adaptive versions of the method are illustrated using the canonical continuous time changepoint problem of inferring the intensity of an inhomogeneous Poisson process, although the method is generally applicable to any changepoint problem. Performance is demonstrated using both conjugate and nonconjugate Bayesian models for the intensity. Appendices to the article are available online, illustrating the method on other models and applications.

Efficient Computation of Bayesian Optimal Discriminating Designs P. 424-433

Holger Dette, Roman Guchenko & Viatcheslav B. Melas

Abstract

An efficient algorithm for the determination of Bayesian optimal discriminating designs for competing regression models is developed, where the main focus is on models with general distributional assumptions beyond the "classical" case of normally distributed homoscedastic errors. For this purpose, we consider a Bayesian version of the Kullback–Leibler (KL). Discretizing the prior distribution leads to local KL-optimal discriminating design problems for a large number of competing models. All currently available methods either require a large amount of computation time or fail to calculate the optimal discriminating design, because they can only deal efficiently with a few model comparisons. In this article, we develop a new algorithm for the determination of Bayesian optimal discriminating designs with respect to the Kullback–Leibler criterion. It is demonstrated that the new algorithm is able to calculate the optimal discriminating designs with reasonable accuracy and computational time in situations where all currently available procedures are either slow or fail.

Adaptive, Delayed-Acceptance MCMC for Targets With Expensive Likelihoods P. 434-444

Chris Sherlock, Andrew Golightly & Daniel A. Henderson

Abstract

When conducting Bayesian inference, delayed-acceptance (DA) Metropolis–Hastings (MH) algorithms and DA pseudo-marginal MH algorithms can be applied when it is computationally expensive to calculate the true posterior or an unbiased estimate thereof, but a computationally cheap approximation is available. A first accept-reject stage is applied, with the cheap approximation substituted for the true posterior in the MH acceptance ratio. Only for those proposals that pass through the first stage is the computationally expensive true posterior (or unbiased estimate thereof) evaluated, with a second accept-reject stage ensuring that detailed balance is satisfied with respect to the intended true posterior. In some scenarios, there is no obvious computationally cheap approximation. A weighted average of previous evaluations of the computationally expensive posterior provides a generic approximation to the posterior. If only the *k*-nearest neighbors have nonzero weights then evaluation of the approximate posterior can be made computationally cheap provided that the points at which the posterior has been evaluated are stored in a multi-dimensional binary tree, known as a KD-tree. The contents of the KD-tree are potentially updated after every computationally intensive evaluation. The resulting adaptive, delayed-acceptance [pseudo-marginal] Metropolis–Hastings algorithm is justified both theoretically and empirically. Guidance on tuning parameters is provided and the methodology is applied to a discretely observed Markov jump process characterizing predator–prey interactions and an

ODE system describing the dynamics of an autoregulatory gene network. Supplementary material for this article is available online.

Divide-and-Conquer With Sequential Monte Carlo P. 445-458

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. A. D. Aston & A. Bouchard-Côté

Abstract

We propose a novel class of Sequential Monte Carlo (SMC) algorithms, appropriate for inference in probabilistic graphical models. This class of algorithms adopts a divide-and-conquer approach based upon an auxiliary tree-structured decomposition of the model of interest, turning the overall inferential task into a collection of recursively solved subproblems. The proposed method is applicable to a broad class of probabilistic graphical models, *including* models with loops. Unlike a standard SMC sampler, the proposed divide-and-conquer SMC employs multiple independent populations of weighted particles, which are resampled, merged, and propagated as the method progresses. We illustrate empirically that this approach can outperform standard methods in terms of the accuracy of the posterior expectation and marginal likelihood approximations. Divide-and-conquer SMC also opens up novel parallel implementation options and the possibility of concentrating the computational effort on the most challenging subproblems. We demonstrate its performance on a Markov random field and on a hierarchical logistic regression problem. Supplementary materials including proofs and additional numerical results are available online.

FFT-Based Fast Computation of Multivariate Kernel Density Estimators With Unconstrained Bandwidth Matrices P. 459-462

Artur Gramacki & Jarosław Gramacki

Abstract

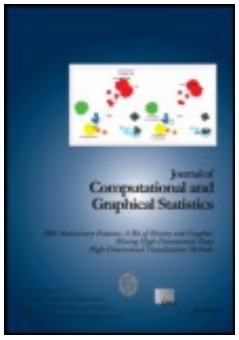
The problem of fast computation of multivariate kernel density estimation (KDE) is still an open research problem. In our view, the existing solutions do not resolve this matter in a satisfactory way. One of the most elegant and efficient approach uses the fast Fourier transform. Unfortunately, the existing FFT-based solution suffers from a serious limitation, as it can accurately operate only with the constrained (i.e., diagonal) multivariate bandwidth matrices. In this article, we describe the problem and give a satisfactory solution. The proposed solution may be successfully used also in other research problems, for example, for the fast computation of the optimal bandwidth for KDE. Supplementary materials for this article are available online.

An Efficient Implementation of the EMICM Algorithm for the Interval Censored NPMLE P. 463-467

Clifford Anderson-Bergman

Abstract

The EMICM algorithm is an established method for computing the interval-censored NPMLE, a generalization of the Kaplan Meier curves for interval censored data. The novel contribution in this work is an efficient implementation, allowing each iteration to be computed in linear time. Using simulated data, it is shown that this new implementation is significantly faster than alternative EMICM implementations or other competing algorithms, allowing for analyses of datasets orders of magnitude larger than previously available.



Journal of computational and graphical statistics, ISSN 1061-8600
Volume 26, number 3 (september 2017)

Abstract

Boxplots are useful displays that convey rough information about the distribution of a variable. Boxplots were designed to be drawn by hand and work best for small datasets, where detailed estimates of tail behavior beyond the quartiles may not be trustworthy. Larger datasets afford more precise estimates of tail behavior, but boxplots do not take advantage of this precision, instead presenting large numbers of extreme, though not unexpected, observations. Letter-value plots address this problem by including more detailed information about the tails using “letter values,” an order statistic defined by Tukey. Boxplots display the first two letter values (the median and quartiles); letter-value plots display further letter values so far as they are reliable estimates of their corresponding quantiles. We illustrate letter-value plots with real data that demonstrate their usefulness for large datasets. All graphics are created using the R package lvplot, and code and data are available in the supplementary materials.

Abstract

The complexity of linear mixed-effects (LME) models means that traditional diagnostics are rendered less effective. This is due to a breakdown of asymptotic results, boundary issues, and visible patterns in residual plots that are introduced by the model fitting process. Some of these issues are well known and adjustments have been proposed. Working with LME models typically requires that the analyst keeps track of all the special circumstances that may arise. In this article, we illustrate a simpler but generally applicable approach to diagnosing LME models. We explain how to use new visual inference methods for these purposes. The approach provides a unified framework for diagnosing LME fits and for model selection. We illustrate the use of this approach on several commonly available datasets. A large-scale Amazon Turk study was used to validate the methods. R code is provided for the analyses. Supplementary materials for this article are available online.

Abstract

Modular programming is a development paradigm that emphasizes self-contained, flexible, and independent pieces of functionality. This practice allows new features to be seamlessly added when desired, and unwanted features to be removed, thus simplifying the software's user interface. The recent rise of web-based software applications has presented new challenges for designing an extensible, modular software system. In this article, we outline a framework for designing such a system, with a focus on reproducibility of the results. We present as a case study a Shiny-based web application called intRo, that allows the user to perform basic data analyses and statistical routines. Finally, we highlight some challenges we

encountered, and how to address them, when combining modular programming concepts with reactive programming as used by Shiny. Supplementary material for this article is available online.

Abstract
In this article, we propose a new framework for matrix factorization based on principal component analysis (PCA) where sparsity is imposed. The structure to impose sparsity is defined in terms of groups of correlated variables found in correlation matrices or maps. The framework is based on three new contributions: an algorithm to identify the groups of variables in correlation maps, a visualization for the resulting groups, and a matrix factorization. Together with a method to compute correlation maps with minimum noise level, referred to as missing-data for exploratory data analysis (MEDA), these three contributions constitute a complete matrix factorization framework. Two real examples are used to illustrate the approach and compare it with PCA, sparse PCA, and structured sparse PCA. Supplementary materials for this article are available online.

Abstract
Contour maps are widely used to display estimates of spatial fields. Instead of showing the estimated field, a contour map only shows a fixed number of contour lines for different levels. However, despite the ubiquitous use of these maps, the uncertainty associated with them has been given a surprisingly small amount of attention. We derive measures of the statistical uncertainty, or quality, of contour maps, and use these to decide an appropriate number of contour lines, which relates to the uncertainty in the estimated spatial field. For practical use in geostatistics and medical imaging, computational methods are constructed, that can be applied to Gaussian Markov random fields, and in particular be used in combination with integrated nested Laplace approximations for latent Gaussian models. The methods are demonstrated on simulated data and an application to temperature estimation is presented.

Abstract
The statistics literature of the past 15 years has established many favorable properties for sparse diminishing-bias regularization: techniques that can roughly be understood as providing estimation under penalty functions spanning the range of concavity between ℓ_0 and ℓ_1 norms. However, lasso ℓ_1-regularized estimation remains the standard tool for industrial Big Data applications because of its minimal computational cost and the presence of easy-to-apply rules for penalty selection. In response, this article proposes a simple new algorithm framework that requires no more computation than a lasso path: the path of one-step estimators (POSE) does ℓ_1 penalized regression estimation on a grid of decreasing penalties, but adapts coefficient-specific weights to decrease as a function of the coefficient estimated in the previous path step. This provides sparse diminishing-bias regularization at no extra cost over the fastest lasso algorithms. Moreover, our gamma lasso implementation of POSE is accompanied by a reliable heuristic for the fit degrees of freedom, so that standard information criteria can be applied in penalty selection. We also provide novel results on the distance between weighted-ℓ_1 and ℓ_0 penalized predictors; this allows us to build intuition about POSE and other diminishing-bias regularization schemes. The methods and results are illustrated in extensive simulations and in application of logistic regression to evaluating the performance of hockey players. Supplementary materials for this article are available online.

Tensor Decomposition With Generalized Lasso Penalties P. 537-546

Oscar Hernan Madrid-Padilla & James Scott

Abstract

We present an approach for penalized tensor decomposition (PTD) that estimates smoothly varying latent factors in multiway data. This generalizes existing work on sparse tensor decomposition and penalized matrix decompositions, in a manner parallel to the generalized lasso for regression and smoothing problems. Our approach presents many nontrivial challenges at the intersection of modeling and computation, which are studied in detail. An efficient coordinate-wise optimization algorithm for PTD is presented, and its convergence properties are characterized. The method is applied both to simulated data and real data on flu hospitalizations in Texas and motion-capture data from video cameras. These results show that our penalized tensor decomposition can offer major improvements on existing methods for analyzing multiway data that exhibit smooth spatial or temporal features.

Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression P. 547-557

Congrui Yi & Jian Huang

Abstract

We propose an algorithm, semismooth Newton coordinate descent (SNCD), for the elastic-net penalized Huber loss regression and quantile regression in high dimensional settings. Unlike existing coordinate descent type algorithms, the SNCD updates a regression coefficient and its corresponding subgradient simultaneously in each iteration. It combines the strengths of the coordinate descent and the semismooth Newton algorithm, and effectively solves the computational challenges posed by dimensionality and nonsmoothness. We establish the convergence properties of the algorithm. In addition, we present an adaptive version of the "strong rule" for screening predictors to gain extra efficiency. Through numerical experiments, we demonstrate that the proposed algorithm is very efficient and scalable to ultrahigh dimensions. We illustrate the application via a real data example. Supplementary materials for this article are available online.

Monitoring Joint Convergence of MCMC Samplers P. 558-568

Douglas VanDerwerken & Scott C. Schmidler

Abstract

We present a diagnostic for monitoring convergence of a Markov chain Monte Carlo (MCMC) sampler to its target distribution. In contrast to popular existing methods, we monitor convergence to the joint target distribution directly rather than a select scalar projection. The method uses a simple nonparametric posterior approximation based on a state-space partition obtained by clustering the pooled draws from multiple chains, and convergence is determined when the estimated posterior probabilities of partition elements under each chain are sufficiently similar. This framework applies to a wide variety of problems, and generalizes directly to non-Euclidean state spaces. Our method also provides approximate high-posterior-density regions, and a characterization of differences between nonconverged chains, all with little additional computational burden. We demonstrate this approach on applications to sampling posterior distributions over Rp, graphs, and partitions. Supplementary materials for this article are available online.

Penalized Nonparametric Scalar-on-Function Regression via Principal Coordinates P. 569-578

Philip T. Reiss, David L. Miller, Pei-Shien Wu & Wen-Yu Hua

Abstract

A number of classical approaches to nonparametric regression have recently been extended to the case of functional predictors. This article introduces a new method of this type, which extends intermediate-rank penalized smoothing to scalar-on-function regression. In the proposed method, which we call *principal coordinate ridge regression*, one regresses the response on leading principal coordinates defined by a relevant distance among the functional predictors, while applying a ridge penalty. Our publicly available implementation, based on generalized additive

modeling software, allows for fast optimal tuning parameter selection and for extensions to multiple functional predictors, exponential family-valued responses, and mixed-effects models. In an application to signature verification data, principal coordinate ridge regression, with dynamic time warping distance used to define the principal coordinates, is shown to outperform a functional generalized linear model. Supplementary materials for this article are available online.

ThrEEBoost: Thresholded Boosting for Variable Selection and Prediction via Estimating Equations P. 579-588

Ben Brown, Christopher J. Miller & Julian Wolfson

Abstract

Most variable selection techniques for high-dimensional models are designed to be used in settings, where observations are independent and completely observed. At the same time, there is a rich literature on approaches to estimation of low-dimensional parameters in the presence of correlation, missingness, measurement error, selection bias, and other characteristics of real data. In this article, we present ThrEEBoost (*Thr*esholded *EEBoost*), a general-purpose variable selection technique which can accommodate such problem characteristics by replacing the gradient of the loss by an estimating function. ThrEEBoost generalizes the previously proposed EEBoost algorithm (Wolfson 2011 Wolfson, J. (2011), "EEBoost: A General Method for Prediction and Variable Selection Based on Estimating Equations," *Journal of the American Statistical Association*, 106, 296–305.[Taylor & Francis Online], [Web of Science ®], [Google Scholar]) by allowing the number of regression coefficients updated at each step to be controlled by a thresholding parameter. Different thresholding parameter values yield different variable selection paths, greatly diversifying the set of models that can be explored; the optimal degree of thresholding can be chosen by cross-validation. ThrEEBoost was evaluated using simulation studies to assess the effects of different threshold values on prediction error, sensitivity, specificity, and the number of iterations to identify minimum prediction error under both sparse and nonsparse true models with correlated continuous outcomes. We show that when the true model is sparse, ThrEEBoost achieves similar prediction error to EEBoost while requiring fewer iterations to locate the set of coefficients yielding the minimum error. When the true model is less sparse, ThrEEBoost has lower prediction error than EEBoost and also finds the point yielding the minimum error more quickly. The technique is illustrated by applying it to the problem of identifying predictors of weight change in a longitudinal nutrition study. Supplementary materials are available online.

Formal Hypothesis Tests for Additive Structure in Random Forests P. 1195-1211

Lucas Mentch & Giles Hooker

Abstract

While statistical learning methods have proved powerful tools for predictive modeling, the black-box nature of the models they produce can severely limit their interpretability and the ability to conduct formal inference. However, the natural structure of ensemble learners like bagged trees and random forests has been shown to admit desirable asymptotic properties when base learners are built with proper subsamples. In this work, we demonstrate that by defining an appropriate grid structure on the covariate space, we may carry out formal hypothesis tests for both variable importance and underlying additive model structure. To our knowledge, these tests represent the first statistical tools for investigating the underlying regression structure in a context such as random forests. We develop notions of total and partial additivity and further demonstrate that testing can be carried out at no additional computational cost by estimating the variance within the process of constructing the ensemble. Furthermore, we propose a novel extension of these testing procedures using random projections to allow for computationally efficient testing procedures that retain high power even when the grid size is much larger than that of the training set.

Finding Singular Features P. 598-609

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli & Larry Wasserman

Abstract

We present a method for finding high density, low-dimensional structures in noisy point clouds. These structures are sets with zero Lebesgue measure with respect to the D-dimensional ambient space and belong to a $d < D$-dimensional space. We call them "singular features." Hunting for singular features corresponds to finding unexpected or unknown structures hidden in point clouds belonging to RD. Our method outputs well-defined sets of dimensions $d < D$. Unlike spectral clustering, the method works well in the presence of noise. We show how to find singular features by first finding ridges in the estimated density, followed by a filtering step based on the eigenvalues of the Hessian of the density. The code for plotting all the figures, with the corresponding plots, and the data files used in the article, are in the folder SupplementaryDocument.zip that can be find at the *http://www.stat.cmu.edu/larry/singular.*

High-Dimensional Multivariate Time Series With Additional Structure P. 610-622

Michael Schweinberger, Sergii Babkin & Katherine B. Ensor

Abstract

High-dimensional multivariate time series are challenging due to the dependent and high-dimensional nature of the data, but in many applications there is additional structure that can be exploited to reduce computing time along with statistical error. We consider high-dimensional vector autoregressive processes with spatial structure, a simple and common form of additional structure. We propose novel high-dimensional methods that take advantage of such structure without making model assumptions about how distance affects dependence. We provide nonasymptotic bounds on the statistical error of parameter estimators in high-dimensional settings and show that the proposed approach reduces the statistical error. An application to air pollution in the USA demonstrates that the estimation approach reduces both computing time and prediction error and gives rise to results that are meaningful from a scientific point of view, in contrast to high-dimensional methods that ignore spatial structure. In practice, these high-dimensional methods can be used to decompose high-dimensional multivariate time series into lower-dimensional multivariate time series that can be studied by other methods in more depth. Supplementary materials for this article are available online.

Regularized Estimation of Piecewise Constant Gaussian Graphical Models: The Group-Fused Graphical Lasso P. 1248-1271

Alexander J. Gibberd & James D. B. Nelson

Abstract

The time-evolving precision matrix of a piecewise-constant Gaussian graphical model encodes the dynamic conditional dependency structure of a multivariate time-series. Traditionally, graphical models are estimated under the assumption that data are drawn identically from a generating distribution. Introducing sparsity and sparse-difference inducing priors, we relax these assumptions and propose a novel regularized M-estimator to jointly estimate both the graph and changepoint structure. The resulting estimator possesses the ability to therefore favor sparse dependency structures and/or smoothly evolving graph structures, as required. Moreover, our approach extends current methods to allow estimation of changepoints that are grouped across multiple dependencies in a system. An efficient algorithm for estimating structure is proposed. We study the empirical recovery properties in a synthetic setting. The qualitative effect of grouped changepoint estimation is then demonstrated by applying the method on a genetic time-course dataset. Supplementary material for this article is available online.

Modeling Time-Varying Effects With Large-Scale Survival Data: An Efficient Quasi-Newton Approach P. 635-645

Kevin He, Yuan Yang, Yanming Li, Ji Zhu & Yi Li

Abstract

Nonproportional hazards models often arise in biomedical studies, as evidenced by a recent national kidney transplant study. During the follow-up, the effects of baseline risk factors, such as patients' comorbidity conditions collected at transplantation, may vary over time. To model such dynamic changes of covariate effects, time-varying survival models have emerged as powerful tools. However, traditional methods of fitting time-varying effects survival model rely on an expansion of the original dataset in a repeated measurement format, which, even with a moderate sample size, leads to an extremely large working dataset. Consequently, the computational burden increases quickly as the sample size grows, and analyses of a large dataset such as our motivating example defy any existing statistical methods and software. We propose a novel application of quasi-Newton iteration method to model time-varying effects in survival analysis. We show that the algorithm converges superlinearly and is computationally efficient for large-scale datasets. We apply the proposed methods, via a stratified procedure, to analyze the national kidney transplant data and study the impact of potential risk factors on post-transplant survival. Supplementary materials for this article are available online.

Approximate Bayesian Computation and Model Assessment for Repulsive Spatial Point Processes P. 646-657

Shinichiro Shirota & Alan E. Gelfand

Abstract

In many applications involving spatial point patterns, we find evidence of inhibition or repulsion. The most commonly used class of models for such settings are the Gibbs point processes. A recent alternative, at least to the statistical community, is the determinantal point process. Here, we examine model fitting and inference for both of these classes of processes in a Bayesian framework. While usual MCMC model fitting can be available, the algorithms are complex and are not always well behaved. We propose using approximate Bayesian computation (ABC) for such fitting. This approach becomes attractive because, though likelihoods are very challenging to work with for these processes, generation of realizations given parameter values is relatively straightforward. As a result, the ABC fitting approach is well-suited for these models. In addition, such simulation makes them well-suited for posterior predictive inference as well as for model assessment. We provide details for all of the above along with some simulation investigation and an illustrative analysis of a point pattern of tree data exhibiting repulsion. R code and datasets are included in the supplementary material.

A Skewed and Heavy-Tailed Latent Random Field Model for Spatial Extremes P. 658-670

Behzad Mahmoudian

Abstract

This article develops Bayesian inference of spatial models with a flexible skew latent structure. Using the multivariate skew-normal distribution of Sahu et al., a valid random field model with stochastic skewing structure is proposed to take into account non-Gaussian features. The skewed spatial model is further improved via scale mixing to accommodate more extreme observations. Finally, the skewed and heavy-tailed random field model is used to describe the parameters of extreme value distributions. Bayesian prediction is done with a well-known Gibbs sampling algorithm, including slice sampling and adaptive simulation techniques. The model performance—as far as the identifiability of the parameters is concerned—is assessed by a simulation study and an analysis of extreme wind speeds across Iran. We conclude that our model provides more satisfactory results according to Bayesian model selection and predictive-based criteria. R code to implement the methods used is available as online supplementary material.

A Generalized Smoother for Linear Ordinary Differential Equations P. 671-681

Michelle Carey, Eugene G. Gath & Kevin Hayes

Abstract

Ordinary differential equations (ODEs) are equalities involving a function and its derivatives that define the evolution of the function over a prespecified domain. The applications of ODEs range from simulation and prediction to control and diagnosis in diverse fields such as engineering, physics, medicine, and finance. Parameter estimation is often required to calibrate these theoretical models to data. While there are many methods for estimating ODE parameters from partially observed data, they are invariably subject to several problems including high computational cost, complex estimation procedures, biased estimates, and large sampling variance. We propose a method that overcomes these issues and produces estimates of the ODE parameters that have less bias, a smaller sampling variance, and a 10-fold improvement in computational efficiency. The package *GenPen* containing the Matlab code to perform the methods described in this article is available online.

Precision Matrix Estimation With ROPE P. 682-694

M. O. Kuismin, J. T. Kemppainen & M. J. Sillanpää

Abstract

It is known that the accuracy of the maximum likelihood-based covariance and precision matrix estimates can be improved by penalized log-likelihood estimation. In this article, we propose a ridge-type operator for the precision matrix estimation, ROPE for short, to maximize a penalized likelihood function where the Frobenius norm is used as the penalty function. We show that there is an explicit closed form representation of a shrinkage estimator for the precision matrix when using a penalized log-likelihood, which is analogous to ridge regression in a regression context. The performance of the proposed method is illustrated by a simulation study and real data applications. Computer code used in the example analyses as well as other supplementary materials for this article are available online.

A Cross-Entropy Approach to the Estimation of Generalized Linear Multilevel Models P. 695-708

Marco Bee, Giuseppe Espa, Diego Giuliani & Flavio Santi

Abstract

In this article, we use the cross-entropy method for noisy optimization for fitting generalized linear multilevel models through maximum likelihood. We propose specifications of the instrumental distributions for positive and bounded parameters that improve the computational performance. We also introduce a new stopping criterion, which has the advantage of being problem-independent. In a second step we find, by means of extensive Monte Carlo experiments, the most suitable values of the input parameters of the algorithm. Finally, we compare the method to the benchmark estimation technique based on numerical integration. The cross-entropy approach turns out to be preferable from both the statistical and the computational point of view. In the last part of the article, the method is used to model the probability of firm exits in the healthcare industry in Italy. Supplemental materials are available online.

Penalized Estimation in Large-Scale Generalized Linear Array Models P. 709-724

Adam Lund, Martin Vincent & Niels Richard Hansen

Abstract

Large-scale generalized linear array models (GLAMs) can be challenging to fit. Computation and storage of its tensor product design matrix can be impossible due to time and memory constraints, and previously considered design matrix free algorithms do not scale well with the dimension of the parameter vector. A new design matrix free algorithm is proposed for computing the penalized maximum likelihood estimate for GLAMs, which, in particular, handles nondifferentiable penalty functions. The proposed algorithm is implemented and available via the R package glamlasso. It combines several ideas—previously considered separately—to obtain sparse estimates while at the same time efficiently exploiting the GLAM structure. In this article, the convergence of the algorithm is treated and the

performance of its implementation is investigated and compared to that of glmnet on simulated as well as real data. It is shown that the computation time for glamlasso scales favorably with the size of the problem when compared to glmnet. Supplementary materials, in the form of R code, data and visualizations of results, are available online.

Link Prediction for Partially Observed Networks P. 725-733

Yunpeng Zhao, Yun-Jhong Wu, Elizaveta Levina & Ji Zhu

Abstract

Link prediction is one of the fundamental problems in network analysis. In many applications, notably in genetics, a partially observed network may not contain any negative examples, that is, edges known for certain to be absent, which creates a difficulty for existing supervised learning approaches. We develop a new method that treats the observed network as a sample of the true network with different sampling rates for positive (true edges) and negative (absent edges) examples. We obtain a relative ranking of potential links by their probabilities, using information on network topology as well as node covariates if available. The method relies on the intuitive assumption that if two pairs of nodes are similar, the probabilities of these pairs forming an edge are also similar. Empirically, the method performs well under many settings, including when the observed network is sparse. We apply the method to a protein–protein interaction network and a school friendship network.

One-Step Generalized Estimating Equations With Large Cluster Sizes P. 734-737

Stuart Lipsitz, Garrett Fitzmaurice, Debajyoti Sinha, Nathanael Hevelone, Jim Hu & Louis L. Nguyen

Abstract

Medical studies increasingly involve a large sample of independent clusters, where the cluster sizes are also large. Our motivating example from the 2010 Nationwide Inpatient Sample (NIS) has 8,001,068 patients and 1049 clusters, with average cluster size of 7627. Consistent parameter estimates can be obtained naively assuming independence, which are inefficient when the intra-cluster correlation (ICC) is high. Efficient generalized estimating equations (GEE) incorporate the ICC and sum all pairs of observations within a cluster when estimating the ICC. For the 2010 NIS, there are 92.6 billion pairs of observations, making summation of pairs computationally prohibitive. We propose a one-step GEE estimator that (1) matches the asymptotic efficiency of the fully iterated GEE; (2) uses a simpler formula to estimate the ICC that avoids summing over all pairs; and (3) completely avoids matrix multiplications and inversions. These three features make the proposed estimator much less computationally intensive, especially with large cluster sizes. A unique contribution of this article is that it expresses the GEE estimating equations incorporating the ICC as a simple sum of vectors and scalars.

Statistically Efficient Thinning of a Markov Chain Sampler P. 738-744

Art B. Owen

Abstract

It is common to subsample Markov chain output to reduce the storage burden. Geyer shows that discarding $k - 1$ out of every k observations will not improve statistical efficiency, as quantified through variance in a given computational budget. That observation is often taken to mean that thinning Markov chain Monte Carlo (MCMC) output cannot improve statistical efficiency. Here, we suppose that it costs one unit of time to advance a Markov chain and then $\theta > 0$ units of time to compute a sampled quantity of interest. For a thinned process, that cost θ is incurred less often, so it can be advanced through more stages. Here, we provide examples to show that thinning will improve statistical efficiency if θ is large and the sample autocorrelations decay slowly enough. If the lag $\ell \geq 1$ autocorrelations of a scalar measurement satisfy $\rho_\ell > \rho_{\ell+1} > 0$, then there is always a $\theta < \infty$ at which thinning becomes more efficient for averages of that scalar. Many sample autocorrelation functions resemble first order AR(1) processes with $\rho_\ell = \rho^{|\ell|}$ for some $-1 < \rho < 1$. For an AR(1) process, it is possible to compute the most efficient subsampling frequency k. The optimal k grows rapidly as ρ increases toward 1. The resulting efficiency gain depends primarily on θ, not ρ. Taking k

= 1 (no thinning) is optimal when $\rho \leq 0$. For $\rho > 0$, it is optimal if and only if $\theta \leq (1 - \rho)^2/(2\rho)$. This efficiency gain never exceeds $1 + \theta$. This article also gives efficiency bounds for autocorrelations bounded between those of two AR(1) processes. Supplementary materials for this article are available online.

Revista de fomento social, ISSN 0015-6043
Volumen 72/2, número 286 (2017)

Resumen
Todo está perdido con la violencia terrorista, nada se logra ni con la destrucción ciega ni con el odio. Todo está perdido con el fanatismo intransigente, todo está perdido con las espirales de violencia, todo está perdido con la complicidad silenciosa, todo está perdido con la ingenuidad y la banalización de los problemas. El bien común es frágil. Es la hora de la responsabilidad, no de la frivolidad; es la hora de la decencia, no de la cobardía; es la hora de la unidad democrática y cívica, no del enfrentamiento estéril.

Resumen
Europa resiste, a pesar de todo. Aunque en un pasado no demasiado lejano un par de referéndum en Francia y en los Países Bajos sirvió para bloquear el proceso de ratificación del proyecto de Constitución europea, éste fue posteriormente reconducido al posterior Tratado de Lisboa, en que se recogieron los aspectos más sustanciales del fracasado proyecto constitucional. Los apocalípticos creen que el llamado ¿brexit¿ es el preludio de la inmediata disolución de la UE, incapaz de hacer frente a la crisis económica y sus consecuencias sociales y ante los desafíos de la nueva realidad mundial y la hegemonía emergente de grandes potencias.
El acontecimiento del ¿brexit¿ ha despertado las conciencias de los europeos. Lo que queda es la imagen de una UE cuajada de perfiles negativos: déficit democrático estructural, altiva burocracia, instituciones bloqueadas, procesos complejos de codecisión y una oleada de populismos de distinto signo que sólo persiguen la destrucción del sistema de la Unión Europea. Para los críticos y los más apocalípticos, desde un punto de vista geoestratégico Europa sería un perdedor neto de la ¿hiperglobalización¿, pues el eje franco¿alemán hace tiempo que dejó de funcionar, lo que unido a su falta de democracia y de transparencia y su incapacidad para enfrentar la principal catástrofe humanitaria de los últimos tiempos, la oleada de refugiados sirios, haría que nos preguntásemos si podemos seguir afirmando que esto es lo normal y previsible, pues la construcción europea siempre ha salido adelante de sucesivas y periódicas crisis con imaginación.Después de describir algunos elementos de la actual crisis de la UE, el editorial plantea la pregunta: ¿han cambiado los valores fundacionales de la UE? Nuestro editorial se inscribe en una reflexión sobre Europa de esta revista iniciada hace años (los títulos de los editoriales figuran en un cuadro final). Hace cinco años nos pronunciábamos sobre la construcción europea a favor de una consolidación, desarrollo y profundización del sistema constitucional europeo de Economía social de mercado. Propugnábamos entonces resocializar el proyecto europeo con nuevas energías, con nueva claridad, con nueva pasión europeísta. Para salir del actual bloqueo, la UE debe liberarse de ese único modelo de capitalismo neoliberal que está precisamente en el origen de la crisis. A continuación de la introducción, abordamos las instituciones europeas y sus resultados, los producidos por sus políticas agrarias y de cohesión regional, muchas veces ignoradas o silenciadas. No silenciamos, por supuesto, los aspectos más problemáticos o negativos, especialmente la insuficiente respuesta europea ante la grave crisis

humanitaria siria que nos pone en evidencia ante el mundo y ante nosotros mismos. Si la capacidad de los gobiernos de las potencias para responder a grandes crisis es tan limitada como se demuestra en estos tiempos, la UE no es una excepción. En el apartado cuarto, sin embargo, presentamos los desafíos con los que se encuentra la UE, con alguna referencia a las políticas españolas desarrolladas a partir del acervo comunitario. El quinto apartado y conclusión nos permite afirmar que la UE es una realidad institucional plenamente consolidada desde hace tiempo, un sistema de economía social de mercado altamente competitiva y solidaria y una aceptable capacidad de insertarse desde un sistema de gobernanza democrática en un mundo hiperglobalizado. Europa puede conjugar una mundialización moderada, la democracia como sistema y la permanencia de los Estados. Nuestra convicción es ésta: la UE está ante una única estrategia realista posible y debe apostar a fondo por ella. Por ello, a pesar de todo ¡Europa!

El caleidoscopio de la integraciónHacia un modelo mixto desde la perspectiva de los migrantes P. 157-201

Alberto Ares Mateos, María Mercedes Fernández García

Resumen

El objetivo principal de este artículo es intentar clarificar y enriquecer el debate de los procesos de cohesión social a través de la propia experiencia de las comunidades migrantes. En este proceso surge la propuesta del Modelo de Integración Mixto (MIM), que estudia los estilos de vida de tres comunidades diversas a través de un enfoque multimétodo basado en una metodología etnográfica y de aculturación a través del consumo. Seguidamente, se exponen los principales hallazgos desde la tipología presentada.

Fundamentos filosóficos de una pedagogía personalista P. 203-253

Tura Pedemonte Feu

Resumen

El ensayo presenta sucesivamente algunas interpelaciones y contribuciones esenciales del pensamiento filosófico europeo del siglo XX que solemos agrupar como personalismo, aunque se recojan otras aportaciones, y busca a partir de ellas la fundamentación de una acción educativa de inspiración personalista. El autor trata especialmente de Paul Ricoeur y de algunos otros pensadores de tradición judía (Espinosa, Rosenzweig, Arendt, Buber, Lévinas). Después de desarrollar las cinco notas que estima más destacables: identidad,narratividad, acción, alteridad y ranscendencia, el ensayo concluye con una invitación a la escuela de inspiración personalista para hacer manifiestos sus presupuestos, concretamente la invitación a la transcendencia desde un respeto profundo por la persona del otro, hasta encontrar el punto justo, aunque difícil, en que la propuesta de lo transcendente, deba concretarse en cada situación que señalará el momento y la forma oportuna de hacerla posible.

Regulación del sector financiero y separación de poderes P. 255-286

Manuel A. Rodríguez Portugués

Resumen

El principio de separación de poderes, cuyo fundamento clásico reside en la garantía de la libertad de los ciudadanos, ha ido mutando conforme a los cambios históricos,desde una visión puramente institucional hasta otra más rica en la que también se tiene en cuenta la idea de equilibrio y contrapeso entre intereses sociales y económicos.En el contexto de ese proceso cobra especial relieve el nacimiento y propagación, dentro de los Estados democráticos constitucionales contemporáneos, de las denominadas autoridades independientes. Sin embargo, la influencia de poderosos intereses económicos y la fagocitación de las instituciones por el denominado Estado de partidos son susceptibles de afectar negativamente al buen funcionamiento de dichos organismos. Para lograr que el principio de separación de poderes continúe respondiendo a su vocación originaria de servicio a la igual libertad de los ciudadanos, el diseño de las autoridades independientes debe respetar una serie de requisitos. Finalmente, en el estudio se aplican dichas ideas a los actuales procesos de reordenación financiera y el organismo encargado de ellos, el Fondo de Reestructuración Ordenada Bancaria (FROB).

Technometrics, ISSN 0040-1706
Volume 59, number 3 (august 2017)

Abstract
Computer experiments with qualitative and quantitative factors occur frequently in various applications in science and engineering. Analysis of such experiments is not yet completely resolved. In this work, we propose an additive Gaussian process model for computer experiments with qualitative and quantitative factors. The proposed method considers an additive correlation structure for qualitative factors, and assumes that the correlation function for each qualitative factor and the correlation function of quantitative factors are multiplicative. It inherits the flexibility of unrestrictive correlation structure for qualitative factors by using the hypersphere decomposition, embracing more flexibility in modeling the complex systems of computer experiments. The merits of the proposed method are illustrated by several numerical examples and a real data application. Supplementary materials for this article are available online.

Abstract
We consider the problem of constructing metamodels for computationally expensive simulation codes; that is, we construct interpolators/predictors of functions values (responses) from a finite collection of evaluations (observations). We use Gaussian process (GP) modeling and kriging, and combine a Bayesian approach, based on a finite set GP models, with the use of localized covariances indexed by the point where the prediction is made. Our approach is not based on postulating a generative model for the unknown function, but by letting the covariance functions depend on the prediction site, it provides enough flexibility to accommodate arbitrary nonstationary observations. Contrary to kriging prediction with plug-in parameter estimates, the resulting Bayesian predictor is constructed explicitly, without requiring any numerical optimization, and locally adjusts the weights given to the different models according to the data variability in each neighborhood. The predictor inherits the smoothness properties of the covariance functions that are used and its superiority over plug-in kriging, sometimes also called empirical-best-linear-unbiased predictor, is illustrated on various examples, including the reconstruction of an oceanographic field over a large region from a small number of observations. Supplementary materials for this article are available online.

Abstract
This article presents a comparison of criteria used to characterize two-level designs for screening purposes. To articulate the relationships among criteria, we focus on 7-factor designs with 16–32 runs and 11-factor designs with 20–48 runs. Screening based on selected designs for each of the run sizes considered is studied with simulation using a forward selection procedure and the Dantzig selector. This article compares Bayesian D-optimal designs,

designs created algorithmically to optimize estimation capacity over various model spaces, and orthogonal designs by estimation-based criteria and simulation. In this way, we furnish both general insights regarding various design approaches, as well as a guide to make a choice among a few final candidate designs. Supplementary materials for this article are available online.

Effective Design-Based Model Selection for Definitive Screening Designs P. 319-329

Bradley Jones & Christopher J. Nachtsheim

Abstract

Since their introduction by Jones and Nachtsheim in 2011 Jones, B., and Nachtsheim, C. J. (2011), "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects," *Journal of Quality Technology*, 43, 1–15.[Web of Science ®], [Google Scholar], definitive screening designs (DSDs) have seen application in fields as diverse as bio-manufacturing, green energy production, and laser etching. One barrier to their routine adoption for screening is due to the difficulties practitioners experience in model selection when both main effects and second-order effects are active. Jones and Nachtsheim showed that for six or more factors, DSDs project to designs in any three factors that can fit a full quadratic model. In addition, they showed that DSDs have high power for detecting all the main effects as well as one two-factor interaction or one quadratic effect as long as the true effects are much larger than the error standard deviation. However, simulation studies of model selection strategies applied to DSDs can disappoint by failing to identify the correct set of active second-order effects when there are more than a few such effects. Standard model selection strategies such as stepwise regression, all-subsets regression, and the Dantzig selector are general tools that do not make use of any structural information about the design. It seems reasonable that a modeling approach that makes use of the known structure of a designed experiment could perform better than more general purpose strategies. This article shows how to take advantage of the special structure of the DSD to obtain the most clear-cut analytical results possible.

An Exact Test of Fit for the Gaussian Linear Model Using Optimal Nonbipartite Matching P. 330-337

Samuel D. Pimentel, Dylan S. Small & Paul R. Rosenbaum

Abstract

Fisher tested the fit of Gaussian linear models using replicated observations. We refine this method by (1) constructing near-replicates using an optimal nonbipartite matching and (2) defining a distance that focuses on predictors important to the model's predictions. Near-replicates may not exist unless the predictor set is low-dimensional; the test addresses dimensionality by betting that model failures involve a subset of predictors important in the old fit. Despite using the old fit to pair observations, the test has exactly its stated level under the null hypothesis. Simulations show the test has reasonable power even when many spurious predictors are present.

Modeling Regression Quantile Process Using Monotone B-Splines P. 338-350

Yuan Yuan, Nan Chen & Shiyu Zhou

Abstract

Quantile regression as an alternative to conditional mean regression (i.e., least-square regression) is widely used in many areas. It can be used to study the covariate effects on the entire response distribution by fitting quantile regression models at multiple different quantiles or even fitting the entire regression quantile process. However, estimating the regression quantile process is inherently difficult because the induced conditional quantile function needs to be monotone at all covariate values. In this article, we proposed a regression quantile process estimation method based on monotone B-splines. The proposed method can easily ensure the validity of the regression quantile process and offers a concise framework for variable selection and adaptive complexity control. We thoroughly investigated the properties of the proposed procedure, both theoretically and numerically. We also used a case study on wind power generation to demonstrate its use and effectiveness in real problems. Supplementary materials for this

article are available online.

Multiple Testing in Regression Models With Applications to Fault Diagnosis in the Big Data Era P. 351-360

Ching-Kang Ing, Tze Leung Lai, Milan Shen, KaWai Tsang & Shu-Hui Yu

Abstract

Motivated by applications to root-cause identification of faults in multistage manufacturing processes that involve a large number of tools or equipment at each stage, we consider multiple testing in regression models whose outputs represent the quality characteristics of a multistage manufacturing process. Because of the large number of input variables that correspond to the tools or equipments used, this falls in the framework of regression modeling in the modern era of big data. On the other hand, with quick fault detection and diagnosis followed by tool rectification, sparsity can be assumed in the regression model. We introduce a new approach to address the multiple testing problem and demonstrate its advantages over existing methods. We also illustrate its performance in an application to semiconductor wafer fabrication that motivated this development. Supplementary materials for this article are available online.

A Geometric Approach to Archetypal Analysis and Nonnegative Matrix Factorization P. 361-370

Anil Damle & Yuekai Sun

Abstract

Archetypal analysis and nonnegative matrix factorization (NMF) are staples in a statistician's toolbox for dimension reduction and exploratory data analysis. We describe a geometric approach to both NMF and archetypal analysis by interpreting both problems as finding extreme points of the data cloud. We also develop and analyze an efficient approach to finding extreme points in high dimensions. For modern massive datasets that are too large to fit on a single machine and must be stored in a distributed setting, our approach makes only a small number of passes over the data. In fact, it is possible to obtain the NMF or perform archetypal analysis with just two passes over the data.

Comparing the Reliability of Related Populations With the Probability of Agreement P. 371-380

Nathaniel T. Stevens & Christine M. Anderson-Cook

Abstract

Combining information from different populations to improve precision, simplify future predictions, or improve underlying understanding of relationships can be advantageous when considering the reliability of several related sets of systems. Using the probability of agreement to help quantify the similarities of populations can help to give a realistic assessment of whether the systems have reliability that are sufficiently similar for practical purposes to be treated as a homogeneous population. The new method is described and illustrated with an example involving two generations of a complex system, where the reliability is modeled using either a logistic or probit regression model. Note that supplementary materials including code, datasets, and added discussion are available online.

A Generalized Quasi-MMSE Controller for Run-to-Run Dynamic Models P. 381-390

Sheng-Tsaing Tseng & Pei-Yu Chen

Abstract

This study proposes a generalized quasi-minimum mean square error (qMMSE) controller for implementing a run-to-run process control where the process input–output relationship follows a general-order dynamical model with added noise. The expression of the process output, the long-term stability conditions and the optimal discount factor of this controller are derived analytically. Furthermore, we use the proposed second-order dynamical model to illustrate the

effects of mis-identification of the process I-O model on the process total mean square error (TMSE). Via a comprehensive simulation study, the model demonstrates that the TMSE may inflate by more than 150% if a second-order dynamical model with moderately large carryover effects is wrongly identified as that of a first-order model. This means that the effects of mis-identification of the process I-O model on the process total mean square error (TMSE) is not negligible for implementing a dynamic run-to-run (RTR) process control. Supplementary materials for this article are available online.

Quantifying Nanoparticle Mixing State to Account for Both Location and Size Effects P. 391-403

Ling Dong, Xiaodong Li, Dan Yu, Hui Zhang, Zhong Zhang, Yanjun Qian & Yu Ding

Abstract

Ripley's *K* function is commonly used to characterize the homogeneity of spatial point distribution. Not surprisingly, it becomes a favored tool in quantifying the nanoparticles mixing state in composite materials, a parameter that material scientists believe is of close relevance to certain properties of the nanoparticle-embedding material. Ripley's *K* function assumes that the spatial points are dimensionless. In reality, the nanoparticles, once mixed in a host material, form clusters or agglomerates of various sizes and shapes. Our analysis shows that using the original *K* function falls short of ranking or distinguishing the homogeneity of nanoparticle mixing. We therefore propose to revise the *K* function to account for both particle location and size effects. We apply the revised function to electron microscopy images of material samples and conduct analysis and comparison of nanoparticle mixing. The analysis shows that the revised function is a better index to quantify the mixing states.