Biblioteca del Instituto de Estadística y Cartografía de Andalucía

Resúmenes de revistas
Enero-febrero 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 mensual 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 en los meses de enero y febrero 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.

AH: Andalucía en la historia, ISSN 1695-1956
Número 55 (enero – marzo 2016)

Dosier. Gitanos: la historia olvidada P. 6-41
Coordinado por: María Sierra

Resumen
El pueblo gitano es uno de los sujetos históricos más olvidados de la historia andaluza, española y europea. Tanto su marginación socioeconómica y política como su cultura tradicionalmente ágrafa han contribuido a la invisibilización de esta comunidad, de la que es un viejo lugar común afirmar que no tiene historia propia. Esto se agrava con los efectos de la acumulación de estereotipos e imágenes generalmente negativas sobre su identidad que producen las sociedades en las que se han insertado históricamente. Este dosier, coordinado por la catedrática de Historia Contemporánea de la Universidad de Sevilla maría Sierra, tiene como objeto mostrar que los gitanos tienen una historia propia y mucho más plural de lo que las visiones más habituales suelen considerar.

El proyecto imperial de Cartago: la "leyenda negra" de los cartagineses P. 42-47
Eduardo Ferrer Albelda

Resumen
Amílcar, Asdrúbal, Aníbal, tres generales de la familia Barca cuyos destinos estuvieron ligados a Hispania. En apenas tres décadas conquistaron el sur y este de la Península Ibérica y proyectaron configurar la provincia más occidental del estado cartaginés. Pero esta pretensión colisionó con los intereses de otra gran potencia, Roma, y el sueño imperial de Cartago se desvaneció de manera violenta.

Piratas cordobeses conquistan Creta P. 48-51
Manuel Huertas

Resumen
En los primeros tiempos de la conquista de Egipto, el califa Omar le preguntó a su general cómo era el mar, a lo que respondió: "El mar es una bestia enorme sobre la que los estúpidos cabalgan como gusanos sobre troncos". Temeroso, el califa dio orden de que ningún musulmán se echase a la mar. pero, pronto tuvieron que vencer el miedo para estar a la par de sus rivales. Sobre todo, si pretendían apoderarse de sus posesiones.

La herencia de la gitana Magdalena de los Reyes: testamento ante notario antes de ser ajusticiada P. 52-53
Ana María Chacón Sánchez-Molina

Resumen
Uno de los primeros testimonios de la presencia de gitanos en Córdoba lo encontramos en el Archivo Histórico Provincial en un protocolo notarial fechado el 4 de agosto de 1596. En este emocionante documento una mujer gitana, presa en la cárcel de Córdoba y condenada a la horca, otorga testamento ante notario y recuerda a su padre, "conde de los gitanos", título con el que se conocía al jefe del clan.

Monstruosidad y medicina: seres monstruosos en academias y colegios de cirugía P. 54-57

Mª Alejandra Flores de la Flor

Resumen

El BOOM teratológico experimentado en la Edad Moderna afectó a numerosos camos del conocimiento: filosofía, historia natural, teología, derecho, etc. Los autores que abordaban este tema se preguntaban qué eran los monstruos y por qué se generaba la monstruosidad y, sobre todo, qué interpetación debían darle y cómo debían enfrentarse a ellos. Las reacciones ante la monstruosidad, desde las más supersticiosas a las más científicas, nos dicen mucho de la mentalidad de una sociedad que respondía de manera contradictoria, a veces opuesta, ante lo que no lograban comprender y llegaban a rechazar: el monstruo.

Chillón, entonces cordobés y ahora manchego: un pueblo en tierra de frontera P. 58-61

Rafael Gil Bautista

Resumen

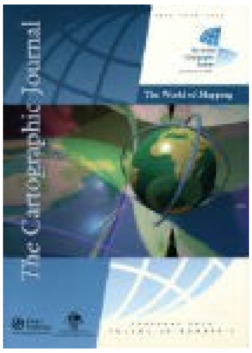
Chillón siempre ha sido un pueblo de frontera. Ubicado en el noroeste cordobés, colindante a La Mancha y Extremadura, tras la reconquista fue un espacio de señorío perteneciendo a los dominios de los Fernández de Córdoba -bien como alcaides de los Donceles, bien con la Casa de Comares- y, más tarde, formando parte de las posesiones de los duques de Medinaceli. Su proximidad a las minas de Almadén fue determinante para que en las últimas décadas del siglo XVIII la Corona lo anexionara a aquellos pozos y fábricas de mercurio. Entonces, como en muchos aspectos también sucede ahora, poco importó la idiosincrasia o los intereses de sus vecinos.

¿Vive la Pepa? Lo que queda de la Constitución de 1812 P. 62-67

Jesús Vallejo

Resumen

Al grito de "¡Viva la Pepa!", los partidarios de la Constitución de Cádiz proclamaron su ahesión a la ley fundamental de 1812. Estuvo en vigor poco tiempo, pero quedó anclada en la memoria de los españoles. Fue muy pronto el símbolo de una libertad bien ganada y digna de ser reconquistada. Los vaivenes políticos consolidaron el mito, que no dejó de agrandarse al compás de los centenarios. Hoy es parte principal del acervo que conforma nuestra identidad nacional, y aún oímos hablar de ella como si sus logros perduraran. Pero ¿vive todavía?¿Están vigentes sus principios y sus preceptos?



Cartographic journal, The, ISSN 0008-7041
Volume 53, number 2 (may 2016)

The Emergence of an Environmental Cartography in Denmark

P. 101-113

Stig Roar Svenningsen

Abstract

Within the history of cartography, relatively little attention has been devoted to the study of the growing body of maps and spatial data focusing on environmental issues. This is rather surprising, considering the importance of this type of cartography in the handling of the complex environmental problems of modern society. This paper analyses the development of thematic maps and spatial data focusing on the terrestrial environment of Danish landscapes. The paper is introduced with a review of the concept of environmental cartography, followed by a historical analysis of the development of environmental mapping in Denmark. Results suggest that there has been a change in the content and aim of environmental cartography in the twentieth century, from an initial focus on mapping potentials for land use improvement and optimization of the economic outputs from engagement with terrestrial ecosystems, to a focus on monitoring ecosystems and regulation of human intervention. Finally, the usefulness of the concept environmental cartography to frame analytical work dealing with the still increasing number of maps produced for environmental purposes within the history of cartography, is evaluated.

Assessing the Planimetric Accuracy of Historical Maps (Sixteenth to Nineteenth Centuries): New Methods and Potential for Coastal Landscape Reconstruction

P. 114-132

Iason Jongepier, Tim Soens, Stijn Temmerman & Tine Missiaen

Abstract

Historical maps are vital tools for landscape reconstruction from the late medieval period onwards. However, the planimetric accuracy of local and regional maps before the nineteenth century is often considered problematic. This paper proposes a method for the evaluation of these maps, through integration in multiple computer programs such as ArcGIS, MapAnalyst and statistical software (SPSS). This method has been tested on a sample of historical maps depicting coastal landscape change in an area at the present-day Dutch-Belgian border (ranging from the local to the supra-regional level and from the sixteenth to the nineteenth centuries), and variations in planimetric accuracy over time have been interpreted. Results point to an exceptionally high accuracy of earlier medium- and large-scale maps – scale being the first determinant of planimetric accuracy – since no significant rise in accuracy over time was found. Notwithstanding this overall accuracy, many maps display pronounced local distortions. However, rather than disqualifying maps for landscape reconstruction, systematic analysis of these distortions can help to facilitate the interpretation of the historical maps and their use for landscape reconstruction. Finally, a method for integrating map accuracies in landscape reconstructions based on multiple maps is proposed and illustrated.

The Town Plans and Sketches of William Stukeley

P. 133-148

Brian Robson & David Bower

Abstract

The eighteenth-century field archaeologist, William Stukeley, travelled widely throughout England to produce the numerous

sketches and plans that illustrated his *Itinerarium Curiosum*. His work has generally not been seen as having made a serious contribution to the cartography or to the portrayal of the towns and landscape of preindustrial England, but the quality of his sketches and the relative accuracy of his town plans are explored here to suggest that this may be too harsh a view.

Chikyû Bankoku Sankai Yochi Zenzu Setsu: The First Japanese World Map with Latitudes and Longitudes and with an Extensive Japanese Explanatory Note P. 149-157

Gabor Lukacs

Abstract

In the early 1780s, Nagakubo Sekisui, the first Japanese scientific geographer, published a world map containing latitudes and longitudes, based on Matteo Ricci's map of 1602. The map and its extensive explanatory text had a considerable impact on the educated classes of the late Edo Period (1603–1868) toward their new vision of the world. We are providing here an analysis of the map and the first complete English translation of Nagakubo Sekisui's most interesting, long explanatory text.

Generalization of the Lambert–Lagrange projection P. 158-165

Sebastian Orihuela

Abstract

The Lagrange projection represents conformally the terrestrial globe within a circle. This is achieved by compressing the latitude and longitude and by applying the new coordinates into the equatorial stereographic projection. The same concept can be generalized to any conformal projection, although the application of this technique to other analytical functions is less known. In this work, the general Lambert–Lagrange projection formula is proposed and the application of the modified coordinates is discussed on projections: stereographic, conformal conic and Gauss–Schreiber. In general, the results are merely a curiosity, except for the case of Gauss–Schreiber, where the use of coordinates with altered scale can be applied in the optimization of conformal projections.

Assessing the Effectiveness and Efficiency of Map Colour for Colour Impairments Using an Eye-tracking Approach P. 166-167

Weihua Dong, Shaobo Zhang, Hua Liao, Zhao Liu, Zhilin Li & Xiaofang Yang

Abstract

Colour impairments influences access to geographical information which is usually represented by colour maps. Three dimensions of colour: Hue, Saturation and Value (HSV), are intuitive and most critical visual variables in map design. In this paper, we specifically focus on colour deficiency of red-green colour impairments. A controlled experiment was designed and conducted to explore how three colour dimensions (HSV) affect the abilities of people with normal colour vision or with red-green colour impairments to distinguish colours in maps. An eye-tracking approach was applied to quantify the accuracy and response time by capturing user eye movements to analyse the effectiveness and efficiency. In this study, we used one section of the administrative map of Hebei Province to test participant responses to area features. Differences of effectiveness and efficiency across normal colour vision and red-green colour impairments were compared. Multiple comparisons among Hue, Saturation and Value were analysed. Results show that for both normal colour vision and red-green colour impairments, Hue is the most differentiable than Saturation and Value. Saturation and Value are at the same level to be differentiated and more difficult to be distinguished. Guidelines of designing maps for both normal colour vision and red-green colour impairments were derived. The results of this study can be helpful to improve the map designs for colour deficiency.

Projection Wizard – An Online Map Projection Selection Tool P. 177-185

Bojan Šavrič, Bernhard Jenny & Helen Jenny

Abstract

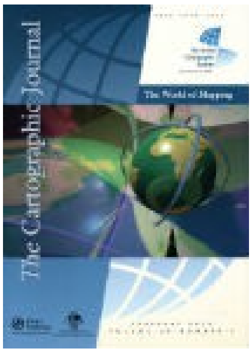
The selection of map projections is difficult and confusing for many. This article introduces Projection Wizard, an online map projection selection tool available at projectionwizard.org that helps mapmakers select projections. The user selects the desired distortion property, and the area to be mapped on an interactive web map. Projection Wizard then proposes a projection, along with projection parameters (such as standard parallels). The tool also creates a preview map with the proposed projection, and provides the corresponding projection code in PROJ.4 format, if applicable. The automated selection process is based on John P. Snyder's selection guideline with a few adjustments. This article discusses the automated selection process, and the map projections suggested. Projection Wizard solves the problem of map projection selection for many applications and helps cartographers and GIS users choose appropriate map projections.

The Use of Mental and Sketch Maps as a Tool to Evaluate Cartography Teaching Effectiveness P. 186-196

Kamil Nieścioruk

Abstract

The paper describes mental maps and their use in teaching process. The survey conducted among students of geodesy and cartography resulted in 124 sketches. They were analysed from the point of view of cartographic methodology and used methods of presentation. The different elements and methods were counted and helped in evaluation of teaching process effectiveness, showing changes in students’ knowledge of certain rules of cartographic language and design and their applications. As the survey was conducted in relation to courses taught, the results are of great value in increasing the quality of cartographic content of these courses and teaching methods.



Cartographic journal, The, ISSN 0008-7041
Volume 53, number 4 (november 2016)

From Mental Maps to GeoParticipation P. 300-307
Jiří Pánek

Abstract

Ever since behavioural geographers started working with place perception; and Peter Gould and Kevin Lynch used mental maps to explore city visualization and spatial preferences, participation has become an integral part of geographical research. Later, when Robert Chambers and others introduced maps into Participatory Rural Appraisal, Participatory GIS and Public Participation GIS were also recognized by quantitative geographers as research methods and visualization tools. In the era of smartphones and global Internet coverage, applications such as FixMyStreet, ArcGIS Online, CartoDB and Maptioannaire allow users to cross the technology gap and become neocartographers without the need for GIS knowledge. GeoParticipation based on using spatial tools in order to involve citizens in community participation can be the future development of Public Participation GIS as it provides an easy-to-use environment and social engagement while creating the feeling of belonging to a certain social group or community. The paper presents a historical review of participatory approaches to the creation of maps, while focusing on the changing role of citizens; from being the objects of geographical research to being the creators of the agenda as well as decision-makers within their communities. Maps were always used as tools of power, but there is a visible shift in the (map) power structures, from maps created by experts and state administration representatives towards maps created by people and their users.

A Shared Perspective for PGIS and VGI P. 308-317
Jeroen Verplanke, Michael K. McCall, Claudia Uberhuaga, Giacomo Rambaldi & Muki Haklay

Abstract

This paper reviews persistent principles of participation processes. On the basis of a review of recent interrogations of the (Public) Participatory Geographic Information Systems (P)PGIS and Volunteered Geographic Information (VGI) approaches, a summary of five prevailing principles in participatory spatial information handling is presented. We investigate these five principles that are common to (P)PGIS and VGI on the basis of a framework of two dimensions that govern the participatory use of spatial information from the perspective of people and society. This framework is presented as a shared perspective of (P)PGIS and VGI and illustrates that, although both share many of these same principles, the ways in which these principles are approached are highly diverse. The paper ends with a future outlook in which we discuss the inter-connected memes of potential technological futures, the signification of localness in 'local spatial knowledge', and the ramifications of ethical tenets by which PGIS and VGI can strengthen each other as two sides of the same coin.

From PGIS to Participatory Deep Mapping and Spatial Storytelling: An Evolving Trajectory in Community Knowledge Representation in GIS P. 318-325
Trevor M. Harris

Abstract

Participatory GIS (PGIS) was borne out of the cauldron of the GIS and Society debates and the social theoretic critique of GIS. The form and practice of PGIS continues to reflect its origins. At its core PGIS remains focused on integrating local

knowledge that is multivalent, equivocal, and often conflictual within a reductionist GIS technology and extensive Spatial Data Infrastructure. Recent conceptual developments in deep mapping and spatial storytelling have the potential to advance the representation of community knowledge through participatory deep mapping. Deep mapping explicitly recognizes that social life is contingent, implicated, and unpredictable. In representing a critical engagement between Geographic Information Science (GISc) and community knowledge and representation, deep mapping potentially challenges the misalignment in representing community knowledge in GIS and in bending geospatial technologies to the needs of communities.

Upside-Down GIS: The Future of Citizen Science and Community Participation P. 326-334
Michelle M. Thompson

Abstract

This article will focus on the changes in time, technology and data that have affected traditional partner relationships using participatory geographic information systems (PGIS). Project development roles of reliance held by the community, and managed by university agents, has shifted from cooperative to, in some cases, complete independence. The modern model of citizen participation includes a resident-planner toolkit with greater access to neighbourhood data and low- to high-tech analytical tools. Many community-led quality of life studies have a limited scope and focus on policy issues that do not serve a larger constituency. Many neighbourhood plans exclude self-reported neighbourhood knowledge and, due to the frequency of municipal reporting cycles, leaves gaps and data mismatch. Given this, the traditional public participation GIS (PPGIS) model may be less data driven due to a more mission-driven resident-led PGIS solution. Planners in practice and in academia have raised levels of concern about data standards, interoperability, reliability, error and metadata. How and why Citizen Science influenced the progression of PPGIS, participation GIS, crowdsourcing and now community-managed data in both theory and practice are provided. This paper will reflect on how top-down strategies to include neighbourhood knowledge are being reframed by the United States Federal Community of Practice. The future of data integration focuses on both the process and products of data development from both the bottom-up and top-down perspectives.

Powering Up: Revisiting Participatory GIS and Empowerment P. 335-340
Jon Corbett, Logan Cochrane & Mark Gill

Abstract

Since 1996, participatory GIS (PGIS) has facilitated avenues through which public participation can occur. One of the ways practitioners articulate social change associated with PGIS interventions has been to qualify success using the term ‘empowerment’. This paper explores the extent to which PGIS academic literature has utilised, defined, measured, and analysed empowerment. This research will demonstrate the degree to which PGIS has, from 1996 to 2014, appropriately and adequately taken into account the causative and direct relationship between a PGIS intervention and empowerment. This article identifies works broadly dealing with PGIS, then searches within that subset of literature for the term ‘empowerment.’ The findings are both quantitatively and qualitatively assessed to explore the trends within the PGIS literature over time and to contextualise the ways in which empowerment has been identified, understood, and articulated. We conclude with a discussion on the extent to which future PGIS research and practice has the ability to disrupt power inequalities.

Facilitating PPGIS Through University Libraries P. 341-347
Rina Ghose & Stephen Appel

Abstract

Equitable access to local geospatial data continues to pose challenges to the knowledge production efforts of marginalized citizen groups. While local government agencies have provided greater access to public data sets through their internet Geographic Information System (GIS) sites, data cannot always be downloaded and used directly by citizens. Past research demonstrates that data sharing at the local level can be a challenging task, mired by legal, institutional, and personal issues. Despite the hype about open data in government, its acceptance and implementation is slow at the local

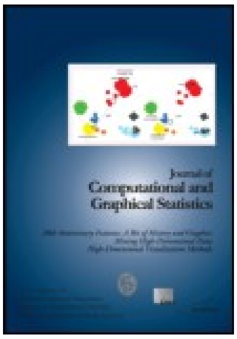
scale. The need for a centralized data repository system at the local scale is thus crucial. This research explores the recent groundbreaking effort to establish a state-wide geospatial portal among the 26 University of Wisconsin (UW) library systems. Through a survey and follow up interviews conducted among public land information professionals in Wisconsin, we find that GIS professionals in local and county governments are open to data sharing through a common geospatial portal. Simultaneously, the efforts to introduce open source GIS software and technical skills through workshops conducted by the library staff demonstrate new ways to facilitate Public Participation GIS (PPGIS). Our research thus demonstrates that university libraries can emerge as an effective model for advancing PPGIS through geoportals, web services, and data and applications in the cloud.

Mapping the Digital Terrain: Towards Indigenous Geographic Information and Spatial Data Quality Indicators for Indigenous Knowledge and Traditional Land-Use Data Collection P. 348-355

Rachel Olson, Jeffrey Hackett & Steven DeRoy

Abstract

Mapping spatial information to represent indigenous knowledge (IK) and rights has been taking place since the early 1970s in various parts of Canada. These mapping initiatives continue to be primarily associated with traditional land-use (TLU) studies and have deep roots in participatory methods that include aspects of participatory geographic information systems (PGIS). In the current context of encroaching industrial developments into indigenous homelands and the strengthening of Indigenous rights within Canadian Supreme Court rulings, the role of mapping TLU information is central. Who is conducting the research, what tools are used, and how this information is shared are all key questions being asked in the Indigenous context. As a result, the quality of spatial data has become a critical part of these engagement processes. This paper focuses on the intersections of new methods of TLU/IK data collection, namely a direct-to-digital approach that seeks to minimize misrepresentation and mistranslations of IK. From these intersections, the authors recognize the need to establish Indigenous-led quality indicators that directly address the introduction of new methods into the TLU/IK field. Indigenous geographic information and spatial data quality indicators will better address the current needs of Indigenous communities in the negotiation of resource developments in their territories, and provide a new path forward for enhancing the use of geospatial technologies in Indigenous communities.



Journal of computational and graphical statistics, ISSN 1061-8600
Volume 25, number 1 (march 2016)

Efficient Implementations of the Generalized Lasso Dual Path Algorithm P. 1-27

Taylor B. Arnold & Ryan J. Tibshirani

Abstract

We consider efficient implementations of the generalized lasso dual path algorithm given by Tibshirani and Taylor in 2011 Tibshirani, R.J., Taylor, J. (2011), The Solution Path of the Generalized Lasso, Annals of Statistics, 39, 1335–1371.[CrossRef], [Web of Science ®]. We first describe a generic approach that covers any penalty matrix *D* and any (full column rank) matrix *X* of predictor variables. We then describe fast implementations for the special cases of trend filtering problems, fused lasso problems, and sparse fused lasso problems, both with *X* = *I* and a general matrix *X*. These specialized implementations offer a considerable improvement over the generic implementation, both in terms of numerical stability and efficiency of the solution path computation. These algorithms are all available for use in the genlasso R package, which can be found in the CRAN repository.

Confidence Areas for Fixed-Effects PCA P. 28-48

Julie Josse, Stefan Wager & François Husson

Abstract

Principal component analysis (PCA) is often used to visualize data when the rows and the columns are both of interest. In such a setting, there is a lack of inferential methods on the PCA output. We study the asymptotic variance of a fixed-effects model for PCA, and propose several approaches to assessing the variability of PCA estimates: a method based on a parametric bootstrap, a new cell-wise jackknife, as well as a computationally cheaper approximation to the jackknife. We visualize the confidence regions by Procrustes rotation. Using a simulation study, we compare the proposed methods and highlight the strengths and drawbacks of each method as we vary the number of rows, the number of columns, and the strength of the relationships between variables.

Case-Specific Random Forests P. 49-65

Ruo Xu, Dan Nettleton & Daniel J. Nordman

Abstract

Random forest (RF) methodology is a nonparametric methodology for prediction problems. A standard way to use RFs includes generating a global RF to predict all test cases of interest. In this article, we propose growing different RFs specific to different test cases, namely case-specific random forests (CSRFs). In contrast to the bagging procedure in the building of standard RFs, the CSRF algorithm takes weighted bootstrap resamples to create individual trees, where we assign large weights to the training cases in close proximity to the test case of interest a priori. Tuning methods are discussed to avoid overfitting issues. Both simulation and real data examples show that the weighted bootstrap resampling used in CSRF construction can improve predictions for specific cases. We also propose a new case-specific variable importance (CSVI) measure as a way to compare the relative predictor variable importance for predicting a particular case. It is possible that the idea of building a predictor case-specifically can be generalized in other areas.

Merging Mixture Components for Clustering Through Pairwise Overlap

P. 66-90

Volodymyr Melnykov

Abstract

Finite mixture models are well known for their flexibility in modeling heterogeneity in data. Model-based clustering is an important application of mixture models, which assumes that each mixture component distribution can adequately model a particular group of data. Unfortunately, when more than one component is needed for each group, the appealing one-to-one correspondence between mixture components and groups of data is ruined and model-based clustering loses its attractive interpretation. Several remedies have been considered in literature. We discuss the most promising recent results obtained in this area and propose a new algorithm that finds partitionings through merging mixture components relying on their pairwise overlap. The proposed technique is illustrated on a popular classification and several synthetic datasets, with excellent results.

Sufficient Dimension Reduction via Distance Covariance

P. 91-104

Wenhui Sheng & Xiangrong Yin

Abstract

We introduce a novel approach to sufficient dimension-reduction problems using distance covariance. Our method requires very mild conditions on the predictors. It estimates the central subspace effectively even when many predictors are categorical or discrete. Our method keeps the model-free advantage without estimating link function. Under regularity conditions, root-*n* consistency and asymptotic normality are established for our estimator. We compare the performance of our method with some existing dimension-reduction methods by simulations and find that our method is very competitive and robust across a number of models. We also analyze the Auto MPG data to demonstrate the efficacy of our method. Supplemental materials for this article are available online.

Assessing the Calibration of High-Dimensional Ensemble Forecasts Using Rank Histograms

P. 105-122

Thordis L. Thorarinsdottir, Michael Scheuerer & Christopher Heinz

Abstract

Any decision-making process that relies on a probabilistic forecast of future events necessarily requires a calibrated forecast. This article proposes new methods for empirically assessing forecast calibration in a multivariate setting where the probabilistic forecast is given by an ensemble of equally probable forecast scenarios. Multivariate properties are mapped to a single dimension through a prerank function and the calibration is subsequently assessed visually through a histogram of the ranks of the observation's preranks. Average ranking assigns a prerank based on the average univariate rank while band depth ranking employs the concept of functional band depth where the centrality of the observation within the forecast ensemble is assessed. Several simulation examples and a case study of temperature forecast trajectories at Berlin Tegel Airport in Germany demonstrate that both multivariate ranking methods can successfully detect various sources of miscalibration and scale efficiently to high-dimensional settings. Supplemental material in form of computer code is available online.

Accounting for Time Series Errors in Partially Linear Model With Single- or Multiple-Runs

P. 123-143

Chunming Zhang, Yu Han & Shengji Jia

Abstract

This article concerns statistical estimation of the partially linear model (PLM) for time course measurements, which are temporally correlated and allow multiple-runs for repeated measurements to enhance experimental accuracy without extending the number of time points within each trial. Such features arise naturally from biomedical data, for example, in brain fMRI, and call for special treatment beyond classical methods in either a purely nonparametric regression model or a PLM with independent errors. We develop a stepwise procedure for estimating the parametric and nonparametric components of the multiple-run PLM and making inference for parameters of interest, adaptive to either single- or multiple-run, in the presence of error temporal dependence. Simulation study and real fMRI data applications illustrate the computational simplicity and effectiveness of the proposed methods. Supplementary material for this article is available online.

Robust Autocorrelation Estimation P. 144-166

Christopher C. Chang & Dimitris N. Politis

Abstract

In this article, we introduce a new class of robust autocorrelation estimators based on interpreting the sample autocorrelation function as a linear regression. We investigate the efficiency and robustness properties of the estimators that result from employing three common robust regression techniques. We discuss the construction of robust autocovariance and positive definite autocorrelation estimates, and their application to AR model fitting. We perform simulation studies with various outlier configurations to compare the different estimators.

Covariance Partition Priors: A Bayesian Approach to Simultaneous Covariance Estimation for Longitudinal Data P. 167-186

J. T. Gaskins & M. J. Daniels

Abstract

The estimation of the covariance matrix is a key concern in the analysis of longitudinal data. When data consist of multiple groups, it is often assumed the covariance matrices are either equal across groups or are completely distinct. We seek methodology to allow borrowing of strength across potentially similar groups to improve estimation. To that end, we introduce a covariance partition prior that proposes a partition of the groups at each measurement time. Groups in the same set of the partition share dependence parameters for the distribution of the current measurement given the preceding ones, and the sequence of partitions is modeled as a Markov chain to encourage similar structure at nearby measurement times. This approach additionally encourages a lower-dimensional structure of the covariance matrices by shrinking the parameters of the Cholesky decomposition toward zero. We demonstrate the performance of our model through two simulation studies and the analysis of data from a depression study. This article includes Supplementary Materials available online.

Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets P. 187-208

Ying Sun & Michael L. Stein

Abstract

For Gaussian process models, likelihood-based methods are often difficult to use with large irregularly spaced spatial datasets, because exact calculations of the likelihood for n observations require $O(n^3)$ operations and $O(n^2)$ memory. Various approximation methods have been developed to address the computational difficulties. In this article, we propose new, unbiased estimating equations (EE) based on score equation approximations that are both computationally and statistically efficient. We replace the inverse covariance matrix that appears in the score equations by a sparse matrix to approximate the quadratic forms, then set the resulting quadratic forms equal to their expected values to obtain unbiased EE. The sparse matrix is constructed by a sparse inverse Cholesky approach to approximate the inverse covariance matrix. The statistical efficiency of the resulting unbiased EE is evaluated both in theory and by numerical studies. Our methods are applied to nearly 90,000 satellite-based measurements of water vapor levels over a region in the Southeast Pacific Ocean.

Nonparametric Estimation for Self-Exciting Point Processes—A Parsimonious Approach P. 209-224

Feng Chen & Peter Hall

Abstract

There is ample evidence that in applications of self-exciting point-process models, the intensity of background events is often far from constant. If a constant background is imposed that assumption can reduce significantly the quality of statistical analysis, in problems as diverse as modeling the after-shocks of earthquakes and the study of ultra-high frequency financial data. Parametric models can be used to alleviate this problem, but they run the risk of distorting inference by misspecifying the nature of the background intensity function. On the other hand, a purely nonparametric approach to analysis leads to problems of identifiability; when a nonparametric approach is taken, not every aspect of the model can be identified from data recorded along a single observed sample path. In this article, we suggest overcoming this difficulty by using an approach based on the principle of parsimony, or Occam's razor. In particular, we suggest taking the point-process intensity to be either a constant or to have maximum differential entropy, in cases where there is not sufficient empirical evidence to suggest that the background intensity function is more complex than those models. This approach is seldom, if ever, used for nonparametric function estimation in other settings, not least because in those cases more data are typically available. However, our "ontological parsimony" argument is appropriate in the context of self-exciting point-process models. Supplementary materials are available online.

Laplace Variational Approximation for Semiparametric Regression in the Presence of Heteroscedastic Errors P. 225-245

Bruce D. Bugbee, F. Jay Breidt & Mark J. van der Woerd

Abstract

Variational approximations provide fast, deterministic alternatives to Markov chain Monte Carlo for Bayesian inference on the parameters of complex, hierarchical models. Variational approximations are often limited in practicality in the absence of conjugate posterior distributions. Recent work has focused on the application of variational methods to models with only partial conjugacy, such as in semiparametric regression with heteroscedastic errors. Here, both the mean and log variance functions are modeled as smooth functions of covariates. For this problem, we derive a mean field variational approximation with an embedded Laplace approximation to account for the nonconjugate structure. Empirical results with simulated and real data show that our approximate method has significant computational advantages over traditional Markov chain Monte Carlo; in this case, a delayed rejection adaptive Metropolis algorithm. The variational approximation is much faster and eliminates the need for tuning parameter selection, achieves good fits for both the mean and log variance functions, and reasonably reflects the posterior uncertainty. We apply the methods to log-intensity data from a small angle X-ray scattering experiment, in which properly accounting for the smooth heteroscedasticity leads to significant improvements in posterior inference for key physical characteristics of an organic molecule.

Joint Modeling of Multiple Network Views P. 246-265

Isabella Gollini & Thomas Brendan Murphy

Abstract

Latent space models (LSM) for network data rely on the basic assumption that each node of the network has an unknown position in a *D*-dimensional Euclidean latent space: generally the smaller the distance between two nodes in the latent space, the greater their probability of being connected. In this article, we propose a variational inference approach to estimate the intractable posterior of the LSM. In many cases, different network views on the same set of nodes are available. It can therefore be useful to build a model able to jointly summarize the information given by all the network views. For this purpose, we introduce the latent space joint model (LSJM) that merges the information given by multiple network views assuming that the probability of a node being connected with other nodes in each network view is explained by a unique latent variable. This model is demonstrated on the analysis of two datasets: an excerpt of 50 girls from "Teenage Friends and Lifestyle Study" data at three time points and the *Saccharomyces cerevisiae* genetic and physical protein–protein interactions. Supplementary materials for this article are available online.

A Pivotal Allocation-Based Algorithm for Solving the Label-Switching Problem in Bayesian Mixture Models

P. 266-283

Han Li & Xiaodan Fan

Abstract

In Bayesian analysis of mixture models, the label-switching problem occurs as a result of the posterior distribution being invariant to any permutation of cluster indices under symmetric priors. To solve this problem, we propose a novel relabeling algorithm and its variants by investigating an approximate posterior distribution of the latent allocation variables instead of dealing with the component parameters directly. We demonstrate that our relabeling algorithm can be formulated in a rigorous framework based on information theory. Under some circumstances, it is shown to resemble the classical Kullback-Leibler relabeling algorithm and include the recently proposed equivalence classes representatives relabeling algorithm as a special case. Using simulation studies and real data examples, we illustrate the efficiency of our algorithm in dealing with various label-switching phenomena. Supplemental materials for this article are available online.

Algorithms for Envelope Estimation

P. 284-300

R. Dennis Cook & Xin Zhang

Abstract

Envelopes were recently proposed as methods for reducing estimative variation in multivariate linear regression. Estimation of an envelope usually involves optimization over Grassmann manifolds. We propose a fast and widely applicable one-dimensional (1D) algorithm for estimating an envelope in general. We reveal an important structural property of envelopes that facilitates our algorithm, and we prove both Fisher consistency and $\sqrt{n}$-consistency of the algorithm. Supplementary materials for this article are available online.

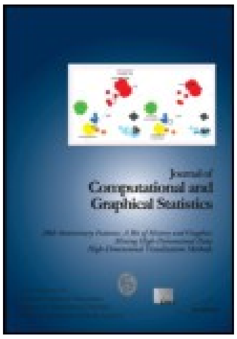
Computational Aspects of Optional Pólya Tree

P. 301-320

Hui Jiang, John Chong Mu, Kun Yang, Chao Du, Luo Lu & Wing Hung Wong

Abstract

Optional Pólya tree (OPT) is a flexible nonparametric Bayesian prior for density estimation. Despite its merits, the computation for OPT inference is challenging. In this article, we present time complexity analysis for OPT inference and propose two algorithmic improvements. The first improvement, named limited-lookahead optional Pólya tree (LL-OPT), aims at accelerating the computation for OPT inference. The second improvement modifies the output of OPT or LL-OPT and produces a continuous piecewise linear density estimate. We demonstrate the performance of these two improvements using simulated and real date examples.



Journal of computational and graphical statistics, ISSN 1061-8600
Volume 25, number 2 (june 2016)

Spline Multiscale Smoothing to Control FDR for Exploring Features of Regression Curves

P. 325-343

Na Li & Xingzhong Xu

Abstract

SiZer (significant zero crossing of the derivatives) is a multiscale smoothing method for exploring trends, maxima, and minima in data. In this article, a regression spline version of SiZer is proposed in a nonparametric regression setting by the fiducial method. The number of knots for spline interpolation is used as the scale parameter of the new SiZer, which controls the smoothness of estimate. In the construction of the new SiZer, multiple testing adjustment is made to control the row-wise false discovery rate (FDR) of SiZer. This adjustment is appealing for exploratory data analysis and has potential to increase the power. A special map is also produced on a continuous scale using *p*-values to assess the significance of features. Simulations and a real data application are carried out to investigate the performance of the proposed SiZer, in which several comparisons with other existing SiZers are presented.

On the Nyström and Column-Sampling Methods for the Approximate Principal Components Analysis of Large Datasets

P. 344-362

Darren Homrighausen & Daniel J. McDonald

Abstract

In this article, we analyze approximate methods for undertaking a principal components analysis (PCA) on large datasets. PCA is a classical dimension reduction method that involves the projection of the data onto the subspace spanned by the leading eigenvectors of the covariance matrix. This projection can be used either for exploratory purposes or as an input for further analysis, for example, regression. If the data have billions of entries or more, the computational and storage requirements for saving and manipulating the design matrix in fast memory are prohibitive. Recently, the Nyström and column-sampling methods have appeared in the numerical linear algebra community for the randomized approximation of the singular value decomposition of large matrices. However, their utility for statistical applications remains unclear. We compare these approximations theoretically by bounding the distance between the induced subspaces and the desired, but computationally infeasible, PCA subspace. Additionally we show empirically, through simulations and a real data example involving a corpus of emails, the trade-off of approximation accuracy and computational complexity.

Exploratory Analysis and Modeling of Stock Returns

P. 363-381

Kimihiro Noguchi, Alexander Aue & Prabir Burman

Abstract

In this article, novel joint semiparametric spline-based modeling of conditional mean and volatility of financial time series is proposed and evaluated on daily stock return data. The modeling includes functions of lagged response variables and time as predictors. The latter can be viewed as a proxy for omitted economic variables contributing to the underlying dynamics. The conditional mean model is additive. The conditional volatility model is multiplicative and linearized with a logarithmic transformation. In addition, a cube-root power transformation is employed to symmetrize the lagged response variables.

Using cubic splines, the model can be written as a multiple linear regression, thereby allowing predictions to be obtained in a simple manner. As outliers are often present in financial data, reliable estimation of the model parameters is achieved by trimmed least-square (TLS) estimation for which a reasonable amount of trimming is suggested. To obtain a parsimonious specification of the model, a new model selection criterion corresponding to TLS is derived. Moreover, the (three-parameter) generalized gamma distribution is identified as suitable for the absolute multiplicative errors and shown to work well for predictions and also for the calculation of quantiles, which is important to determine the value at risk. All model choices are motivated by a detailed analysis of IBM, HP, and SAP daily returns. The prediction performance is compared to the classical generalized autoregressive conditional heteroskedasticity (GARCH) and asymmetric power GARCH (APGARCH) models as well as to a nonstationary time-trend volatility model. The results suggest that the proposed model may possess a high predictive power for future conditional volatility.

Penalized Fast Subset Scanning

P. 382-404

Skyler Speakman, Sriram Somanchi, Edward McFowland III & Daniel B. Neill

Abstract

We present the penalized fast subset scan (PFSS), a new and general framework for scalable and accurate pattern detection. PFSS enables exact and efficient identification of the most anomalous subsets of the data, as measured by a likelihood ratio scan statistic. However, PFSS also allows incorporation of prior information about each data element's probability of inclusion, which was not previously possible within the subset scan framework. PFSS builds on two main results: first, we prove that a large class of likelihood ratio statistics satisfy a property that allows additional, element-specific penalty terms to be included while maintaining efficient computation. Second, we prove that the penalized statistic can be maximized exactly by evaluating only $O(N)$ subsets. As a concrete example of the PFSS framework, we incorporate "soft" constraints on spatial proximity into the spatial event detection task, enabling more accurate detection of irregularly shaped spatial clusters of varying sparsity. To do so, we develop a distance-based penalty function that rewards spatial compactness and penalizes spatially dispersed clusters. This approach was evaluated on the task of detecting simulated anthrax bio-attacks, using real-world Emergency Department data from a major U.S. city. PFSS demonstrated increased detection power and spatial accuracy as compared to competing methods while maintaining efficient computation.

Parameter Expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data

P. 405-425

Lizhen Xu, Radu V. Craiu, Lei Sun & Andrew D. Paterson

Abstract

Motivated by genetic association studies of pleiotropy, we propose a Bayesian latent variable approach to jointly study multiple outcomes. The models studied here can incorporate both continuous and binary responses, and can account for serial and cluster correlations. We consider Bayesian estimation for the model parameters, and we develop a novel MCMC algorithm that builds upon hierarchical centering and parameter expansion techniques to efficiently sample from the posterior distribution. We evaluate the proposed method via extensive simulations and demonstrate its utility with an application to an association study of various complication outcomes related to Type 1 diabetes.

Online Variational Bayes Inference for High-Dimensional Correlated Data

P. 426-444

Sylvie (Tchumtchoua) Kabisa, David B. Dunson & Jeffrey S. Morris

Abstract

High-dimensional data with hundreds of thousands of observations are becoming commonplace in many disciplines. The analysis of such data poses many computational challenges, especially when the observations are correlated over time and/or across space. In this article, we propose flexible hierarchical regression models for analyzing such data that accommodate serial and/or spatial correlation. We address the computational challenges involved in fitting these models by adopting an approximate inference framework. We develop an online variational Bayes algorithm that works by incrementally reading the data into memory one portion at a time. The performance of the method is assessed through simulation studies.

The methodology is applied to analyze signal intensity in MRI images of subjects with knee osteoarthritis, using data from the Osteoarthritis Initiative.

s-CorrPlot: An Interactive Scatterplot for Exploring Correlation P. 445-463

Sean McKenna, Miriah Meyer, Christopher Gregg & Samuel Gerber

Abstract

The degree of correlation between variables is used in many data analysis applications as a key measure of interdependence. The most common techniques for exploratory analysis of pairwise correlation in multivariate datasets, like scatterplot matrices and clustered heatmaps, however, do not scale well to large datasets, either computationally or visually. We present a new visualization that is capable of encoding pairwise correlation between hundreds of thousands variables, called the s-CorrPlot. The s-CorrPlot encodes correlation spatially between variables as points on scatterplot using the geometric structure underlying Pearson's correlation. Furthermore, we extend the s-CorrPlot with interactive techniques that enable animation of the scatterplot to new projections of the correlation space, as illustrated in the companion video in supplementary materials. We provide the s-CorrPlot as an open-source R package and validate its effectiveness through a variety of methods including a case study with a biology collaborator.

Estimation Stability With Cross-Validation (ESCV) P. 464-492

Chinghway Lim & Bin Yu

Abstract

Cross-validation (CV) is often used to select the regularization parameter in high-dimensional problems. However, when applied to the sparse modeling method Lasso, CV leads to models that are unstable in high-dimensions, and consequently not suited for reliable interpretation. In this article, we propose a model-free criterion ESCV based on a new *estimation stability* (ES) metric and CV. Our proposed ESCV finds a smaller and locally ES-optimal model smaller than the CV choice so that it fits the data and also enjoys estimation stability property. We demonstrate that ESCV is an effective alternative to CV at a similar easily parallelizable computational cost. In particular, we compare the two approaches with respect to several performance measures when applied to the Lasso on both simulated and real datasets. For dependent predictors common in practice, our main finding is that ESCV cuts down false positive rates often by a large margin, while sacrificing little of true positive rates. ESCV usually outperforms CV in terms of parameter estimation while giving similar performance as CV in terms of prediction. For the two real datasets from neuroscience and cell biology, the models found by ESCV are less than half of the model sizes by CV, but preserves CV's predictive performance and corroborates with subject knowledge and independent work. We also discuss some regularization parameter alignment issues that come up in both approaches.

Sparse Penalized Forward Selection for Support Vector Classification P. 493-514

Subhashis Ghosal, Bradley Turnbull, Hao Helen Zhang & Wook Yeon Hwang

Abstract

We propose a new binary classification and variable selection technique especially designed for high-dimensional predictors. Among many predictors, typically, only a small fraction of them have significant impact on prediction. In such a situation, more interpretable models with better prediction accuracy can be obtained by variable selection along with classification. By adding an ℓ_1-type penalty to the loss function, common classification methods such as logistic regression or support vector machines (SVM) can perform variable selection. Existing penalized SVM methods all attempt to jointly solve all the parameters involved in the penalization problem altogether. When data dimension is very high, the joint optimization problem is very complex and involves a lot of memory allocation. In this article, we propose a new penalized forward search technique that can reduce high-dimensional optimization problems to one-dimensional optimization by iterating the selection steps. The new algorithm can be regarded as a forward selection version of the penalized SVM and its variants. The advantage of optimizing in one dimension is that the location of the optimum solution can be obtained with intelligent search by exploiting convexity and a piecewise linear or quadratic structure of the criterion function. In each step, the predictor that is most able to predict the outcome is chosen in the model. The search is then repeatedly used in an iterative fashion until convergence occurs. Comparison of our new classification rule with ℓ_1-SVM and other common methods show very promising performance, in that the proposed method leads to much leaner models without compromising misclassification rates, particularly for high-dimensional predictors.

Bayesian Variable Selection on Model Spaces Constrained by Heredity Conditions P. 515-535

Daniel Taylor-Rodriguez, Andrew Womack & Nikolay Bliznyuk

Abstract

This article investigates Bayesian variable selection when there is a hierarchical dependence structure on the inclusion of predictors in the model. In particular, we study the type of dependence found in polynomial response surfaces of orders two and higher, whose model spaces are required to satisfy weak or strong heredity conditions. These conditions restrict the inclusion of higher-order terms depending upon the inclusion of lower-order parent terms. We develop classes of priors on the model space, investigate their theoretical and finite sample properties, and provide a Metropolis–Hastings algorithm for searching the space of models. The tools proposed allow fast and thorough exploration of model spaces that account for hierarchical polynomial structure in the predictors and provide control of the inclusion of false positives in high posterior probability models.

Fast Hamiltonian Monte Carlo Using GPU Computing P. 536-548

Andrew L. Beam, Sujit K. Ghosh & Jon Doyle

Abstract

In recent years, the Hamiltonian Monte Carlo (HMC) algorithm has been found to work more efficiently compared to other popular Markov chain Monte Carlo (MCMC) methods (such as random walk Metropolis–Hastings) in generating samples from a high-dimensional probability distribution. HMC has proven more efficient in terms of mixing rates and effective sample size than previous MCMC techniques, but still may not be sufficiently fast for particularly large problems. The use of GPUs promises to push HMC even further greatly increasing the utility of the algorithm. By expressing the computationally intensive portions of HMC (the evaluations of the probability kernel and its gradient) in terms of linear or element-wise operations, HMC can be made highly amenable to the use of graphics processing units (GPUs). A multinomial regression example demonstrates the promise of GPU-based HMC sampling. Using GPU-based memory objects to perform the entire HMC simulation, most of the latency penalties associated with transferring data from main to GPU memory can be avoided. Thus, the proposed computational framework may appear conceptually very simple, but has the potential to be applied to a wide class of hierarchical models relying on HMC sampling. Models whose posterior density and corresponding gradients can be reduced to linear or element-wise operations are amenable to significant speed ups through the use of GPUs. Analyses of datasets that were previously intractable for fully Bayesian approaches due to the prohibitively high computational cost are now feasible using the proposed framework.

Direction-Projection-Permutation for High-Dimensional Hypothesis Tests P. 549-569

Susan Wei, Chihoon Lee, Lindsay Wichers & J. S. Marron

Abstract

High-dimensional low sample size (HDLSS) data are becoming increasingly common in statistical applications. When the data can be partitioned into two classes, a basic task is to construct a classifier that can assign objects to the correct class. Binary linear classifiers have been shown to be especially useful in HDLSS settings and preferable to more complicated classifiers because of their ease of interpretability. We propose a computational tool called direction-projection-permutation (DiProPerm), which rigorously assesses whether a binary linear classifier is detecting statistically significant differences between two high-dimensional distributions. The basic idea behind DiProPerm involves working directly with the one-dimensional projections of the data induced by binary linear classifier. Theoretical properties of DiProPerm are studied under the HDLSS asymptotic regime whereby dimension diverges to infinity while sample size remains fixed. We show that certain variations of DiProPerm are consistent and that consistency is a nontrivial property of tests in the HDLSS asymptotic regime. The practical utility of DiProPerm is demonstrated on HDLSS gene expression microarray datasets. Finally, an empirical power study is conducted comparing DiProPerm to several alternative two-sample HDLSS tests to understand the advantages and disadvantages of each method.

Soft Null Hypotheses: A Case Study of Image Enhancement Detection in Brain Lesions

P. 570-588

Haochang Shou, Russell T. Shinohara, Han Liu, Daniel S. Reich & Ciprian M. Crainiceanu

Abstract

This work is motivated by a study of a population of multiple sclerosis (MS) patients using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) to identify active brain lesions. At each visit, a contrast agent is administered intravenously to a subject and a series of images are acquired to reveal the location and activity of MS lesions within the brain. Our goal is to identify the enhancing lesion locations at the subject level and lesion enhancement patterns at the population level. We analyze a total of 20 subjects scanned at 63 visits (∼30Gb), the largest population of such clinical brain images. After addressing the computational challenges, we propose possible solutions to the difficult problem of transforming a qualitative scientific null hypothesis, such as “this voxel does not enhance,” to a well-defined and numerically testable null hypothesis based on the existing data. We call such procedure “soft null” hypothesis testing as opposed to the standard “hard null” hypothesis testing. This problem is fundamentally different from: (1) finding testing statistics when a quantitative null hypothesis is given; (2) clustering using a mixture distribution; or (3) setting a reasonable threshold with a parametric null assumption.

Bayesian Ising Graphical Model for Variable Selection

P. 589-605

Zaili Fang & Inyoung Kim

Abstract

In this article, we propose a new Bayesian variable selection (BVS) approach via the graphical model and the Ising model, which we refer to as the “Bayesian Ising graphical model” (BIGM). The BIGM is developed by showing that the BVS problem based on the linear regression model can be considered as a complete graph and described by an Ising model with random interactions. There are several advantages of our BIGM: it is easy to (i) employ the single-site updating and cluster updating algorithm, both of which are suitable for problems with small sample sizes and a larger number of variables, (ii) extend this approach to nonparametric regression models, and (iii) incorporate graphical prior information. In our BIGM, the interactions are determined by the linear model coefficients, so we systematically study the performance of different scale normal mixture priors for the model coefficients by adopting the global-local shrinkage strategy. Our results indicate that the best prior for the model coefficients in terms of variable selection should place substantial weight on small, nonzero shrinkage. The methods are illustrated with simulated and real data.

Tweedie’s Compound Poisson Model With Grouped Elastic Net

P. 606-625

Wei Qian, Yi Yang & Hui Zou

Abstract

Wei Qian is Assistant Professor, School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623 (E-mail: *wxqsma@rit.edu*). Yi Yang is Assistant Professor, Department of Mathematics and Statistics, McGill University, Canada (E-mail: *yi.yang6@mcgill.ca*) Hui Zou is Professor of Statistics, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *zouxx019@umn.edu*).

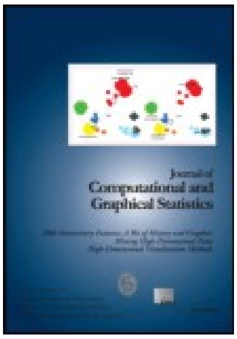
Tweedie’s compound Poisson model is a popular method to model data with probability mass at zero and nonnegative, highly right-skewed distribution. Motivated by wide applications of the Tweedie model in various fields such as actuarial science, we investigate the grouped elastic net method for the Tweedie model in the context of the generalized linear model. To efficiently compute the estimation coefficients, we devise a two-layer algorithm that embeds the blockwise majorization descent method into an iteratively reweighted least square strategy. Integrated with the strong rule, the proposed algorithm is implemented in an easy-to-use R package HDtweedie, and is shown to compute the whole solution path very efficiently. Simulations are conducted to study the variable selection and model fitting performance of various lasso methods for the Tweedie model. The modeling applications in risk segmentation of insurance business are illustrated by analysis of an auto insurance claim dataset.

Parallel Variational Bayes for Large Datasets With an Application to Generalized Linear Mixed Models P. 626-646

Minh-Ngoc Tran, David J. Nott, Anthony Y. C. Kuk & Robert Kohn

Abstract

The article develops a hybrid variational Bayes (VB) algorithm that combines the mean-field and stochastic linear regression fixed-form VB methods. The new estimation algorithm can be used to approximate any posterior without relying on conjugate priors. We propose a divide and recombine strategy for the analysis of large datasets, which partitions a large dataset into smaller subsets and then combines the variational distributions that have been learned in parallel on each separate subset using the hybrid VB algorithm. We also describe an efficient model selection strategy using cross-validation, which is straightforward to implement as a by-product of the parallel run. The proposed method is applied to fitting generalized linear mixed models. The computational efficiency of the parallel and hybrid VB algorithm is demonstrated on several simulated and real datasets.



Journal of computational and graphical statistics, ISSN 1061-8600
Volume 25, number 3 (september 2016)

Power-Conditional-Expected Priors: Using *g*-Priors With Random Imaginary Data for Variable Selection P. 647-664
Dimitris Fouskakis & Ioannis Ntzoufras

Abstract
The Zellner's *g*-prior and its recent hierarchical extensions are the most popular default prior choices in the Bayesian variable selection context. These prior setups can be expressed as power-priors with fixed set of imaginary data. In this article, we borrow ideas from the power-expected-posterior (PEP) priors to introduce, under the *g*-prior approach, an extra hierarchical level that accounts for the imaginary data uncertainty. For normal regression variable selection problems, the resulting power-conditional-expected-posterior (PCEP) prior is a conjugate normal-inverse gamma prior that provides a consistent variable selection procedure and gives support to more parsimonious models than the ones supported using the *g*-prior and the hyper-*g* prior for finite samples. Detailed illustrations and comparisons of the variable selection procedures using the proposed method, the *g*-prior, and the hyper-*g* prior are provided using both simulated and real data examples.

Bayesian Sparse Group Selection P. 665-683
Ray-Bing Chen, Chi-Hsiang Chu, Shinsheng Yuan & Ying Nian Wu

Abstract
This article proposes a Bayesian approach for the sparse group selection problem in the regression model. In this problem, the variables are partitioned into different groups. It is assumed that only a small number of groups are active for explaining the response variable, and it is further assumed that within each active group only a small number of variables are active. We adopt a Bayesian hierarchical formulation, where each candidate group is associated with a binary variable indicating whether the group is active or not. Within each group, each candidate variable is also associated with a binary indicator, too. Thus, the sparse group selection problem can be solved by sampling from the posterior distribution of the two layers of indicator variables. We adopt a group-wise Gibbs sampler for posterior sampling. We demonstrate the proposed method by simulation studies as well as real examples. The simulation results show that the proposed method performs better than the sparse group Lasso in terms of selecting the active groups as well as identifying the active variables within the selected groups.

A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo P. 684-700
Lei Gong & James M. Flegal

Abstract
A current challenge for many Bayesian analyses is determining when to terminate high-dimensional Markov chain Monte Carlo simulations. To this end, we propose using an automated sequential stopping procedure that terminates the simulation when the computational uncertainty is small relative to the posterior uncertainty. Further, we show this stopping rule is equivalent to stopping when the effective sample size is sufficiently large. Such a stopping rule has previously been shown to work well in settings with posteriors of moderate dimension. In this article, we illustrate its utility in high-dimensional

simulations while overcoming some current computational issues. As examples, we consider two complex Bayesian analyses on spatially and temporally correlated datasets. The first involves a dynamic space-time model on weather station data and the second a spatial variable selection model on fMRI brain imaging data. Our results show the sequential stopping rule is easy to implement, provides uncertainty estimates, and performs well in high-dimensional settings.

Toward Automatic Model Comparison: An Adaptive Sequential Monte Carlo Approach

P. 701-726

Yan Zhou, Adam M. Johansen & John A.D. Aston

Abstract

Model comparison for the purposes of selection, averaging, and validation is a problem found throughout statistics. Within the Bayesian paradigm, these problems all require the calculation of the posterior probabilities of models within a particular class. Substantial progress has been made in recent years, but difficulties remain in the implementation of existing schemes. This article presents adaptive sequential Monte Carlo (SMC) sampling strategies to characterize the posterior distribution of a collection of models, as well as the parameters of those models. Both a simple product estimator and a combination of SMC and a path sampling estimator are considered and existing theoretical results are extended to include the path sampling variant. A novel approach to the automatic specification of distributions within SMC algorithms is presented and shown to outperform the state of the art in this area. The performance of the proposed strategies is demonstrated via an extensive empirical study. Comparisons with state-of-the-art algorithms show that the proposed algorithms are always competitive, and often substantially superior to alternative techniques, at equal computational cost and considerably less application-specific implementation effort.

Zero Expectile Processes and Bayesian Spatial Regression

P. 727-747

Anandamayee Majumdar & Debashis Paul

Abstract

We introduce new classes of stationary spatial processes with asymmetric, sub-Gaussian marginal distributions using the idea of expectiles. We derive theoretical properties of the proposed processes. Moreover, we use the proposed spatial processes to formulate a spatial regression model for point-referenced data where the spatially correlated errors have skewed marginal distribution. We introduce a Bayesian computational procedure for model fitting and inference for this class of spatial regression models. We compare the performance of the proposed method with the traditional Gaussian process-based spatial regression through simulation studies and by applying it to a dataset on air pollution in California.

Bayesian Ensemble Trees (BET) for Clustering and Prediction in Heterogeneous Data

P. 748-761

Leo L. Duan, John P. Clancy & Rhonda D. Szczesniak

Abstract

We propose a novel “tree-averaging” model that uses the ensemble of classification and regression trees (CART). Each constituent tree is estimated with a subset of similar data. We treat this grouping of subsets as Bayesian ensemble trees (BET) and model them as a Dirichlet process. We show that BET determines the optimal number of trees by adapting to the data heterogeneity. Compared with the other ensemble methods, BET requires much fewer trees and shows equivalent prediction accuracy using weighted averaging. Moreover, each tree in BET provides variable selection criterion and interpretation for each subset. We developed an efficient estimating procedure with improved estimation strategies in both CART and mixture models. We demonstrate these advantages of BET with simulations and illustrate the approach with a real-world data example involving regression of lung function measurements obtained from patients with cystic fibrosis.

GPU-Powered Shotgun Stochastic Search for Dirichlet Process Mixtures of Gaussian Graphical Models

P. 762-788

Chiranjit Mukherjee & Abel Rodriguez

Abstract

Gaussian graphical models (GGMs) are popular for modeling high-dimensional multivariate data with sparse conditional dependencies. A mixture of GGMs extends this model to the more realistic scenario where observations come from a heterogenous population composed of a small number of homogeneous subgroups. In this article, we present a novel stochastic search algorithm for finding the posterior mode of high-dimensional Dirichlet process mixtures of decomposable GGMs. Further, we investigate how to harness the massive thread-parallelization capabilities of graphical processing units to accelerate computation. The computational advantages of our algorithms are demonstrated with various simulated data examples in which we compare our stochastic search with a Markov chain Monte Carlo (MCMC) algorithm in moderate dimensional data examples. These experiments show that our stochastic search largely outperforms the MCMC algorithm in terms of computing-times and in terms of the quality of the posterior mode discovered. Finally, we analyze a gene expression dataset in which MCMC algorithms are too slow to be practically useful.

Parallel Resampling in the Particle Filter

P. 789-805

Lawrence M. Murray, Anthony Lee & Pierre E. Jacob

Abstract

Modern parallel computing devices, such as the graphics processing unit (GPU), have gained significant traction in scientific and statistical computing. They are particularly well-suited to data-parallel algorithms such as the particle filter, or more generally sequential Monte Carlo (SMC), which are increasingly used in statistical inference. SMC methods carry a set of weighted *particles* through repeated propagation, weighting, and resampling steps. The propagation and weighting steps are straightforward to parallelize, as they require only independent operations on each particle. The resampling step is more difficult, as standard schemes require a collective operation, such as a sum, across particle weights. Focusing on this resampling step, we analyze two alternative schemes that do not involve a collective operation (Metropolis and rejection resamplers), and compare them to standard schemes (multinomial, stratified, and systematic resamplers). We find that, in certain circumstances, the alternative resamplers can perform significantly faster on a GPU, and to a lesser extent on a CPU, than the standard approaches. Moreover, in single precision, the standard approaches are numerically biased for upward of hundreds of thousands of particles, while the alternatives are not. This is particularly important given greater single- than double-precision throughput on modern devices, and the consequent temptation to use single precision with a greater number of particles. Finally, we provide auxiliary functions useful for implementation, such as for the permutation of ancestry vectors to enable in-place propagation.

Reinforced Angle-Based Multicategory Support Vector Machines

P. 806-825

Chong Zhang, Yufeng Liu, Junhui Wang & Hongtu Zhu

Abstract

The support vector machine (SVM) is a very popular classification tool with many successful applications. It was originally designed for binary problems with desirable theoretical properties. Although there exist various multicategory SVM (MSVM) extensions in the literature, some challenges remain. In particular, most existing MSVMs make use of k classification functions for a k-class problem, and the corresponding optimization problems are typically handled by existing quadratic programming solvers. In this article, we propose a new group of MSVMs, namely, the reinforced angle-based MSVMs (RAMSVMs), using an angle-based prediction rule with $k-1$ functions directly. We prove that RAMSVMs can enjoy Fisher consistency. Moreover, we show that the RAMSVM can be implemented using the very efficient coordinate descent algorithm on its dual problem. Numerical experiments demonstrate that our method is highly competitive in terms of computational speed, as well as classification prediction performance.

Sparse Distance Weighted Discrimination

P. 826-838

Boxiang Wang & Hui Zou

Abstract

Distance weighted discrimination (DWD) was originally proposed to handle the data piling issue in the support vector machine. In this article, we consider the sparse penalized DWD for high-dimensional classification. The state-of-the-art algorithm for solving the standard DWD is based on second-order cone programming, however such an algorithm does not work well for the sparse penalized DWD with high-dimensional data. To overcome the challenging computation difficulty, we develop a very efficient algorithm to compute the solution path of the sparse DWD at a given fine grid of regularization parameters. We implement the algorithm in a publicly available R package sdwd. We conduct extensive numerical experiments to demonstrate the computational efficiency and classification performance of our method.

Fast and Flexible ADMM Algorithms for Trend Filtering

P. 839-858

Aaditya Ramdas & Ryan J. Tibshirani

Abstract

This article presents a fast and robust algorithm for trend filtering, a recently developed nonparametric regression tool. It has been shown that, for estimating functions whose derivatives are of bounded variation, trend filtering achieves the minimax optimal error rate, while other popular methods like smoothing splines and kernels do not. Standing in the way of a more widespread practical adoption, however, is a lack of scalable and numerically stable algorithms for fitting trend filtering estimates. This article presents a highly efficient, specialized alternating direction method of multipliers (ADMM) routine for trend filtering. Our algorithm is competitive with the specialized interior point methods that are currently in use, and yet is far more numerically robust. Furthermore, the proposed ADMM implementation is very simple, and, importantly, it is flexible enough to extend to many interesting related problems, such as sparse trend filtering and isotonic trend filtering. Software for our method is freely available, in both the C and R languages.

Supervised Sparse and Functional Principal Component Analysis

P. 859-878

Gen Li, Haipeng Shen & Jianhua Z. Huang

Abstract

Principal component analysis (PCA) is an important tool for dimension reduction in multivariate analysis. Regularized PCA methods, such as sparse PCA and functional PCA, have been developed to incorporate special features in many real applications. Sometimes additional variables (referred to as supervision) are measured on the same set of samples, which can potentially drive low-rank structures of the primary data of interest. Classical PCA methods cannot make use of such supervision data. In this article, we propose a supervised sparse and functional principal component (SupSFPC) framework that can incorporate supervision information to recover underlying structures that are more interpretable. The framework unifies and generalizes several existing methods and flexibly adapts to the practical scenarios at hand. The SupSFPC model is formulated in a hierarchical fashion using latent variables. We develop an efficient modified expectation-maximization (EM) algorithm for parameter estimation. We also implement fast data-driven procedures for tuning parameter selection. Our comprehensive simulation and real data examples demonstrate the advantages of SupSFPC.

SHAH: SHape-Adaptive Haar Wavelets for Image Processing

P. 879-898

Piotr Fryzlewicz & Catherine Timmermans

Abstract

We propose the shape-adaptive Haar (SHAH) transform for images, which results in an orthonormal, adaptive decomposition of the image into Haar-wavelet-like components, arranged hierarchically according to decreasing importance, whose shapes reflect the features present in the image. The decomposition is as sparse as it can be for piecewise-constant images. It is performed via a stepwise bottom-up algorithm with quadratic computational complexity; however, nearly linear variants also exist. SHAH is rapidly invertible. We show how to use SHAH for image denoising. Having performed the SHAH transform, the coefficients are hard- or soft-thresholded, and the inverse transform taken. The SHAH image denoising algorithm compares favorably to the state of the art for piecewise-constant images. A clear asset of the methodology is its very general scope: it can be used with any images or more generally with any data that can be represented as graphs or networks.

Fast Nonparametric Density-Based Clustering of Large Datasets Using a Stochastic Approximation Mean-Shift Algorithm

P. 899-916

Ollivier Hyrien & Andrea Baran

Abstract

Mean-shift is an iterative procedure often used as a nonparametric clustering algorithm that defines clusters based on the modal regions of a density function. The algorithm is conceptually appealing and makes assumptions neither about the shape of the clusters nor about their number. However, with a complexity of $O(n^2)$ per iteration, it does not scale well to large datasets. We propose a novel algorithm which performs density-based clustering much quicker than mean shift, yet delivering virtually identical results. This algorithm combines subsampling and a stochastic approximation procedure to achieve a potential complexity of $O(n)$ at each step. Its convergence is established. Its performances are evaluated using simulations and applications to image segmentation, where the algorithm was tens or hundreds of times faster than mean shift, yet causing negligible amounts of clustering errors. The algorithm can be combined with existing approaches to further accelerate clustering.

Testing for Hermite Rank in Gaussian Subordination Processes

P. 917-934

Jan Beran, Sven Möhrle & Sucharita Ghosh

Abstract

Statistical inference for time series with long-range dependence is often based on the assumption of Gaussian subordination $X_t = G(Z_t)$. Although the Hermite rank m of G plays an essential role for statistical inference in these situations, the question of estimating m or of testing hypotheses about the Hermite rank has not been addressed in the literature. In this article, a method is introduced for testing H_0: $m = 1$ against H_1: $m > 1$. This allows for deciding whether inference based on the usual assumption of $m = 1$ is appropriate. Simulations and data examples illustrate the method.

Parsimonious and Efficient Likelihood Composition by Gibbs Sampling

P. 935-953

Davide Ferrari, Guoqi Qian & Tane Hunter

Abstract

The traditional maximum likelihood estimator (MLE) is often of limited use in complex high-dimensional data due to the intractability of the underlying likelihood function. Maximum composite likelihood estimation (McLE) avoids full likelihood specification by combining a number of partial likelihood objects depending on small data subsets, thus enabling inference for complex data. A fundamental difficulty in making the McLE approach practicable is the selection from numerous candidate likelihood objects for constructing the composite likelihood function. In this article, we propose a flexible Gibbs sampling scheme for optimal selection of sub-likelihood components. The sampled composite likelihood functions are shown to converge to the one maximally informative on the unknown parameters in equilibrium, since sub-likelihood objects are chosen with probability depending on the variance of the corresponding McLE. A penalized version of our method generates sparse likelihoods with a relatively small number of components when the data complexity is intense. Our algorithms are illustrated through numerical examples on simulated data as well as real genotype single nucleotide polymorphism (SNP) data from a case–control study.

RAPTT: An Exact Two-Sample Test in High Dimensions Using Random Projections P. 954-970

Radhendushka Srivastava, Ping Li & David Ruppert

Abstract

In high dimensions, the classical Hotelling's T^2 test tends to have low power or becomes undefined due to singularity of the sample covariance matrix. In this article, this problem is overcome by projecting the data matrix onto lower dimensional subspaces through multiplication by random matrices. We propose RAPTT (RAndom Projection T^2-Test), an exact test for equality of means of two normal populations based on projected lower dimensional data. RAPTT does not require any constraints on the dimension of the data or the sample size. A simulation study indicates that in high dimensions the power of this test is often greater than that of competing tests. The advantages of RAPTT are illustrated on a high-dimensional gene expression dataset involving the discrimination of tumor and normal colon tissues.

Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes P. 971-979

Gregory B. Gloor, Jean M. Macklaim & Andrew D. Fernandes

Abstract

Displaying the component-wise between-group differences high-dimensional datasets is problematic because widely used plots such as Bland–Altman and Volcano plots do not show what they are colloquially *believed* to show. Thus, it is difficult for the experimentalist to grasp why the between-group difference of one component is "significant" while that of another component is not. Here, we propose a type of "Effect Plot" that displays between-group differences in relation to respective underlying variability for every component of a high-dimensional dataset. We use synthetic data to show that such a plot captures the essence of what determines "significance" for between-group differences in each component, and provide guidance in the interpretation of the plot. Supplementary online materials contain the code and data for this article and include simple R functions to produce an effect plot from suitable datasets.

Technometrics, ISSN 0040-1706
Volume 58, number 1 (february 2016)

Modeling an Augmented Lagrangian for Blackbox Constrained Optimization

P. 1-11

Robert B. Gramacy, Genetha A. Gray, Sébastien Le Digabel, Herbert K. H. Lee, Pritam Ranjan,
Garth Wells & Stefan M. Wild

Abstract

Constrained blackbox optimization is a difficult problem, with most approaches coming from the mathematical programming literature. The statistical literature is sparse, especially in addressing problems with nontrivial constraints. This situation is unfortunate because statistical methods have many attractive properties: global scope, handling noisy objectives, sensitivity analysis, and so forth. To narrow that gap, we propose a combination of response surface modeling, expected improvement, and the augmented Lagrangian numerical optimization framework. This hybrid approach allows the statistical model to think globally and the augmented Lagrangian to act locally. We focus on problems where the constraints are the primary bottleneck, requiring expensive simulation to evaluate and substantial modeling effort to map out. In that context, our hybridization presents a simple yet effective solution that allows existing objective-oriented statistical approaches, like those based on Gaussian process surrogates and expected improvement heuristics, to be applied to the constrained setting with minor modification. This work is motivated by a challenging, real-data benchmark problem from hydrology where, even with a simple linear objective function, learning a nontrivial valid region complicates the search for a global minimum.

Rejoinder

P. 26-29

Robert B. Gramacy, Genetha A. Gray, Sébastien Le Digabel, Herbert K. H. Lee, Pritam Ranjan,
Garth Wells & Stefan M. Wild

Abstract

We are grateful for the many insightful comments provided by the discussants. One team politely pointed out oversights in our literature review and the subsequent omission of a formidable comparator. Another made an important clarification about when a more aggressive variation (the so-called NoMax) would perform poorly. A third team offered enhancements to the framework, including a derivation of closed-form expressions and a more aggressive updating scheme; these enhancements were supported by an empirical study comparing new alternatives with old. The last team suggested hybridizing the statistical augmented Lagrangian (AL) method with modern stochastic search. Here we present our responses to these contributions and detail some improvements made to our own implementations in light of them. We conclude with some thoughts on statistical optimization using surrogate modeling and open-source software.

Sequential Design of Computer Experiments for the Assessment of Fetus Exposure to Electromagnetic Fields

P. 30-42

Marjorie Jala, Céline Levy-Leduc, Éric Moulines, Emmanuelle Conil & Joe Wiart

Abstract

In this article, we describe four sequential sampling strategies for estimating the quantile of $Y = f(X)$, where X has a known distribution in and f is a deterministic unknown, expensive-to-evaluate real-valued function. These approaches all consist in modeling *f* as a sample of a well-chosen Gaussian process and aim at estimating the quantile by using as few evaluations of *f* as possible. The different methodologies are first compared through various numerical experiments. Then, in the framework of the ANR-JST FETUS project, we apply our strategies to a real example corresponding to the exposure of a Japanese pregnant-woman model and her 26-week-old fetus to a plane wave. Finally, we compare our methodologies on a simplified geometric model designed for modeling the fetus exposure to plane waves.

Optimizing Two-Level Supersaturated Designs Using Swarm Intelligence Techniques

P. 43-49

Frederick Kin Hing Phoa, Ray-Bing Chen, Weichung Wang & Weng Kee Wong

Abstract

Supersaturated designs (SSDs) are often used to reduce the number of experimental runs in screening experiments with a large number of factors. As more factors are used in the study, the search for an optimal SSD becomes increasingly challenging because of the large number of feasible selection of factor level settings. This article tackles this discrete optimization problem via an algorithm based on swarm intelligence. Using the commonly used $E(s^2)$ criterion as an illustrative example, we propose an algorithm to find $E(s^2)$-optimal SSDs by showing that they attain the theoretical lower bounds found in previous literature. We show that our algorithm consistently produces SSDs that are at least as efficient as those from the traditional CP exchange method in terms of computational effort, frequency of finding the $E(s^2)$-optimal SSD, and also has good potential for finding D_3-, D_4-, and D_5-optimal SSDs.

Sliced Orthogonal Array-Based Latin Hypercube Designs

P. 50-61

Youngdeok Hwang, Xu He & Peter Z.G. Qian

Abstract

We propose an approach for constructing a new type of design, called a sliced orthogonal array-based Latin hypercube design. This approach exploits a slicing structure of orthogonal arrays with strength two and makes use of sliced random permutations. Such a design achieves one- and two-dimensional uniformity and can be divided into smaller Latin hypercube designs with one-dimensional uniformity. Sampling properties of the proposed designs are derived. Examples are given for illustrating the construction method and corroborating the derived theoretical results. Potential applications of the constructed designs include uncertainty quantification of computer models, computer models with qualitative and quantitative factors, cross-validation and efficient allocation of computing resources.

A Bayesian Perspective on the Analysis of Unreplicated Factorial Experiments Using Potential Outcomes

P. 62-73

Valeria Espinosa, Tirthankar Dasgupta & Donald B. Rubin

Abstract

Unreplicated factorial designs have been widely used in scientific and industrial settings, when it is important to distinguish “active” or real factorial effects from “inactive” or noise factorial effects used to estimate residual or “error” terms. We propose a new approach to screen for active factorial effects from such experiments that uses the potential outcomes framework and is based on sequential posterior predictive model checks. One advantage of the proposed method is its ability to broaden the standard definition of active effects and to link their definition to the population of interest. Another important aspect of this approach is its conceptual connection to Fisherian randomization tests. Extensive simulation studies are conducted, which demonstrate the superiority of the proposed approach over existing ones in the situations considered.

Blocking Schemes for Definitive Screening Designs

P. 74-83

Bradley Jones & Christopher J. Nachtsheim

Abstract

In earlier work, Jones and Nachtsheim proposed a new class of screening designs called definitive screening designs. As originally presented, these designs are three-level designs for quantitative factors that provide estimates of main effects that are unbiased by any second-order effect and require only one more than twice as many runs as there are factors. Definitive screening designs avoid direct confounding of any pair of second-order effects, and, for designs that have more than five factors, project to efficient response surface designs for any two or three factors. Recently, Jones and Nachtsheim expanded the applicability of these designs by showing how to include any number of two-level categorical factors. However, methods for blocking definitive screening designs have not been addressed. In this article we develop orthogonal blocking schemes for definitive screening designs. We separately consider the cases where all of the factors are quantitative and where there is a mix of quantitative and two-level qualitative factors. The schemes are quite flexible in that the numbers of blocks may vary from two to the number of factors, and block sizes need not be equal. We provide blocking schemes for both fixed and random blocks.

Abstract

Markov chain Monte Carlo approaches have been widely used for Bayesian inference. The drawback of these methods is that they can be computationally prohibitive especially when complex models are analyzed. In such cases, variational methods may provide an efficient and attractive alternative. However, the variational methods reported to date are applicable to relatively simple models and most are based on a factorized approximation to the posterior distribution. Here, we propose a variational approach that is capable of handling models that consist of a system of differential-algebraic equations and whose posterior approximation can be represented by a multivariate distribution. Under the proposed approach, the solution of the variational inference problem is decomposed into three steps: a maximum a posteriori optimization, which is facilitated by using an orthogonal collocation approach, a preprocessing step, which is based on the estimation of the eigenvectors of the posterior covariance matrix, and an expected propagation optimization problem. To tackle multivariate integration, we employ quadratures derived from the Smolyak rule (sparse grids). Examples are reported to elucidate the advantages and limitations of the proposed methodology. The results are compared to the solutions obtained from a Markov chain Monte Carlo approach. It is demonstrated that significant computational savings can be gained using the proposed approach.

Abstract

This article develops methods to estimate the tail and full distribution of the lengths of the 0-intervals in a continuous time stationary ergodic stochastic process that takes the values 0 and 1 in alternating intervals. The setting is that each of many such 0–1 processes has been observed during a short time window. Thus, the observed 0-intervals could be noncensored, right-censored, left-censored, or doubly-censored, and the lengths of 0-intervals that are ongoing at the beginning of the observation window have a length-biased distribution. We exhibit parametric conditional maximum likelihood estimators for the full distribution, develop maximum likelihood tail estimation methods based on a semiparametric generalized Pareto model, and propose goodness-of-fit plots. Finite sample properties are studied by simulation, and asymptotic normality is established for the most important case. The methods are applied to estimation of the length of off-road glances in the 100-car study, a big naturalistic driving experiment.

Abstract
Most research of run-to-run process control has been based on single-input and single-output processes with static input–output relationships. In practice, many complicated semiconductor manufacturing processes have multiple-input and multiple-output (MIMO) variables. In addition, the effects of previous process input recipes and output responses on the current outputs might be carried over for several process periods. Under these circumstances, using conventional controllers usually results in unsatisfactory performance. To overcome this, a complicated process could be viewed as dynamic MIMO systems with added general process disturbance and this article proposes a dynamic-process multivariate exponentially weighted moving average (MEWMA) controller to adjust those processes. The long-term stability conditions of the proposed controller are derived analytically. Furthermore, by minimizing the total mean square error (TMSE) of the process outputs, the optimal discount matrix of the proposed controller under vector IMA(1,□1) disturbance is derived. Finally, to highlight the contribution of the proposed controller, we also conduct a comprehensive simulation study to compare the control performance of the proposed controller with that of the single MEWMA and self-tuning controllers. On average, the results demonstrate that the proposed controller outperforms the other two controllers with a TMSE reduction about 32% and 43%, respectively.

Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection P. 116-126

J. L. Wadsworth

Abstract
To model the tail of a distribution, one has to define the threshold above or below which an extreme value model produces a suitable fit. Parameter stability plots, whereby one plots maximum likelihood estimates of supposedly threshold-independent parameters against threshold, form one of the main tools for threshold selection by practitioners, principally due to their simplicity. However, one repeated criticism of these plots is their lack of interpretability, with pointwise confidence intervals being strongly dependent across the range of thresholds. In this article, we exploit the independent-increments structure of maximum likelihood estimators to produce complementary plots with greater interpretability, and suggest a simple likelihood-based procedure that allows for automated threshold selection.

Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation P. 127-137

Ying Sun, Huixia J. Wang & Montserrat Fuentes

Abstract
Quantile functions are important in characterizing the entire probability distribution of a random variable, especially when the tail of a skewed distribution is of interest. This article introduces new quantile function estimators for spatial and temporal data with a fused adaptive Lasso penalty to accommodate the dependence in space and time. This method penalizes the difference among neighboring quantiles, hence it is desirable for applications with features ordered in time or space without replicated observations. The theoretical properties are investigated and the performances of the proposed methods are evaluated by simulations. The proposed method is applied to particulate matter (PM) data from the Community Multiscale Air Quality (CMAQ) model to characterize the upper quantiles, which are crucial for studying spatial association between PM concentrations and adverse human health effects.

Short-Term Wind Speed Forecast Using Measurements From Multiple Turbines in A Wind Farm P. 138-147

Arash Pourhabib, Jianhua Z. Huang & Yu Ding

Abstract
Turbine operations in a wind farm benefit from an understanding of the near-ground behavior of wind speeds. This article describes a probabilistic spatial-temporal model for analyzing local wind fields. Our model is constructed based on measurements taken from a large number of turbines in a wind farm, as opposed to aggregating the data into a single time-series. The model incorporates both temporal and spatial characteristics of wind speed data: in addition to using a time epoch mechanism to model temporal nonstationarity, our model identifies an informative neighborhood of turbines that are spatially related, and consequently, constructs an ensemble-like predictor using the data associated with the neighboring turbines. Using actual wind data measured at 200 wind turbines in a wind farm, we found that the two modeling elements benefit short-term wind speed forecasts. We also investigate the use of regime switching to account for the effect of wind direction and the use of geostrophic wind to account for the effects of meteorologic factors other than wind. These at best provide a small performance boost to speed forecast.

Technometrics, ISSN 0040-1706
Volume 58, number 1 (february 2016)

Technometrics

A Journal of Statistics for the Physical, Chemical and Engineering Sciences

ASQ and American Statistical Association

Technometrics, ISSN 0040-1706
Volume 58, number 2 (may 2016)

Scaled Predictor Envelopes and Partial Least-Squares Regression P. 155-165

R. Dennis Cook & Zhihua Su

Abstract

Partial least squares (PLS) is a widely used method for prediction in applied statistics, especially in chemometrics applications. However, PLS is not invariant or equivariant under scale transformations of the predictors, which tends to limit its scope to regressions in which the predictors are measured in the same or similar units. Cook, Helland, and Su (2013 Cook, R.D., Helland, I.S., Su, Z. (2013), Envelopes and Partial Least Squares Regression, Journal of the Royal Statistical Society, Series B,75, 851–877.[CrossRef]) built a connection between nascent envelope methodology and PLS, allowing PLS to be addressed in a traditional likelihood-based framework. In this article, we use the connection between PLS and envelopes to develop a new method—scaled predictor envelopes (SPE)—that incorporates predictor scaling into PLS-type applications. By estimating the appropriate scales, the SPE estimators can offer efficiency gains beyond those given by PLS, and further reduce prediction errors. Simulations and an example are given to support the theoretical claims.

Bayesian Additive Regression Tree Calibration of Complex High-Dimensional Computer Models P. 166-179

M. T. Pratola & D. M. Higdon

Abstract

Complex natural phenomena are increasingly investigated by the use of a complex computer simulator. To leverage the advantages of simulators, observational data need to be incorporated in a probabilistic framework so that uncertainties can be quantified. A popular framework for such experiments is the statistical computer model calibration experiment. A limitation often encountered in current statistical approaches for such experiments is the difficulty in modeling high-dimensional observational datasets and simulator outputs as well as high-dimensional inputs. As the complexity of simulators seems to only grow, this challenge will continue unabated. In this article, we develop a Bayesian statistical calibration approach that is ideally suited for such challenging calibration problems. Our approach leverages recent ideas from Bayesian additive regression Tree models to construct a random basis representation of the simulator outputs and observational data. The approach can flexibly handle high-dimensional datasets, high-dimensional simulator inputs, and calibration parameters while quantifying important sources of uncertainty in the resulting inference. We demonstrate our methodology on a CO_2 emissions rate calibration problem, and on a complex simulator of subterranean radionuclide dispersion, which simulates the spatial–temporal diffusion of radionuclides released during nuclear bomb tests at the Nevada Test Site. Supplementary computer code and datasets are available online.

Monotonic Quantile Regression With Bernstein Polynomials for Stochastic Simulation P. 180-190

Matthias H. Y. Tan

Abstract

Quantile regression is an important tool to determine the quality level of service, product, and operation systems via stochastic simulation. It is frequently known that the quantiles of the output distribution are monotonic functions of certain inputs to the simulation model. Because there is typically high variability in estimation of tail quantiles, it can be valuable to incorporate this information in quantile modeling. However, the existing literature on monotone quantile regression with multiple inputs is sparse. In this article, we propose a class of monotonic regression models, which consists of functional analysis of variance (FANOVA) decomposition components modeled with Bernstein polynomial bases for estimating quantiles as a function of multiple inputs. The polynomial degrees of the bases for the model and the FANOVA components included in the model are selected by a greedy algorithm. Real examples demonstrate the advantages of incorporating the monotonicity assumption in quantile regression and the good performance of the proposed methodology for estimating quantiles. Supplementary materials for this article are available online.

A Change-Point Approach for Phase-I Analysis in Multivariate Profile Monitoring and Diagnosis P. 191-204

Kamran Paynabar, Changliang Zou & Peihua Qiu

Abstract

Process monitoring and fault diagnosis using profile data remains an important and challenging problem in statistical process control (SPC). Although the analysis of profile data has been extensively studied in the SPC literature, the challenges associated with monitoring and diagnosis of multichannel (multiple) nonlinear profiles are yet to be addressed. Motivated by an application in multioperation forging processes, we propose a new modeling, monitoring, and diagnosis framework for phase-I analysis of multichannel profiles. The proposed framework is developed under the assumption that different profile channels have similar structure so that we can gain strength by borrowing information from all channels. The multidimensional functional principal component analysis is incorporated into change-point models to construct monitoring statistics. Simulation results show that the proposed approach has good performance in identifying change-points in various situations compared with some existing methods. The codes for implementing the proposed procedure are available in the supplementary material.

Bayesian Detection of Changepoints in Finite-State Markov Chains for Multiple Sequences P. 205-213

Petter Arnesen, Tracy Holsclaw & Padhraic Smyth

Abstract

We consider the analysis of sets of categorical sequences consisting of piecewise homogenous Markov segments. The sequences are assumed to be governed by a common underlying process with segments occurring in the same order for each sequence. Segments are defined by a set of unobserved changepoints where the positions and number of changepoints can vary from sequence to sequence. We propose a Bayesian framework for analyzing such data, placing priors on the locations of the changepoints and on the transition matrices and using Markov chain Monte Carlo (MCMC) techniques to obtain posterior samples given the data. Experimental results using simulated data illustrate how the methodology can be used for inference of posterior distributions for parameters and changepoints, as well as the ability to handle considerable variability in the locations of the changepoints across different sequences. We also investigate the application of the approach to sequential data from an application involving monsoonal rainfall patterns. Supplementary materials for this article are available online.

Bayesian Inference for a New Class of Distributions on Equivalence Classes of Three-Dimensional Orientations With Applications to Materials Science P. 214-224

Chuanlong Du, Daniel J. Nordman & Stephen B. Vardeman

Abstract

Experiments in materials science investigating cubic crystalline structures often collect data which are in truth *equivalence classes* of crystallographically symmetric orientations. These intend to represent how lattice structures of particles are orientated relative to a reference coordinate system. Motivated by a materials science application, we formulate parametric probability models for “unlabeled orientation data.” This amounts to developing models on equivalence classes of three-dimensional rotations. We use a flexible existing model class for random rotations (called uniform-axis-random-spin models) to induce probability distributions on the equivalence classes of rotations. We develop one-sample Bayesian inference for the parameters in these models, and compare this methodology to some likelihood-based approaches. We also contrast the new parametric analysis of unlabeled orientation data with other analyses that proceed as if the data have been preprocessed into honest orientation data. Supplementary materials for this article are available online.

Prior-Free Probabilistic Prediction of Future Observations P. 225-235

Ryan Martin & Rama T. Lingham

Abstract

Prediction of future observations is a fundamental problem in statistics. Here we present a general approach based on the recently developed inferential model (IM) framework. We employ an IM-based technique to marginalize out the unknown parameters, yielding prior-free probabilistic prediction of future observables. Verifiable sufficient conditions are given for validity of our IM for prediction, and a variety of examples demonstrate the proposed method’s performance. Thanks to its generality and ease of implementation, we expect that our IM-based method for prediction will be a useful tool for practitioners. Supplementary materials for this article are available online.

Quantifying Uncertainty in Lumber Grading and Strength Prediction: A Bayesian Approach P. 236-243

Samuel W.K. Wong, Conroy Lum, Lang Wu & James V. Zidek

Abstract

This article presents a joint distribution for the strength of a randomly selected piece of structural lumber and its observable characteristics. In the process of lumber strength testing, these characteristics are ascertained under strict grading protocols, as they have the potential to be strength reducing. However, for practical reasons, only a few such selected characteristics among the many present, are recorded. We present a data-generating mechanism that reflects the uncertainties resulting from the grading protocol. A Bayesian approach is then adopted for model fitting and construction of a predictive distribution for strength that accounts for the unrecorded characteristics. The method is validated on simulated examples, and then applied on a sample of specimens tested for bending and tensile strength. Use of the predictive distribution is demonstrated, and insights gained into the grading process are described. Details of the lumber testing experiments can be found in the online supplementary materials.

Optimum Allocation Rule for Accelerated Degradation Tests With a Class of Exponential-Dispersion Degradation Models P. 244-254

Sheng-Tsaing Tseng & I-Chen Lee

Abstract

Optimum allocation problem in accelerated degradation tests (ADTs) is an important task for reliability analysts. Several researchers have attempted to address this decision problem, but their results have been based only on specific degradation models. Therefore, they lack a unified approach toward general degradation models. This study proposes a class of exponential dispersion (ED) degradation models to overcome this difficulty. Assuming that the underlying degradation path comes from the ED class, we analytically derive the optimum allocation rules (by minimizing the asymptotic variance of the estimated q quantile of product's lifetime) for two-level and three-level ADT allocation problems whether the testing stress levels are prefixed or not. For a three-level allocation problem, we show that all test units should be allocated into two out of three stresses, depending on certain specific conditions. Two examples are used to illustrate the proposed procedure. Furthermore, the penalties of using nonoptimum allocation rules are also addressed. This study demonstrates that a three-level compromise plan with small proportion allocation in the middle stress, in general, is a good strategy for ADT allocation. Supplementary materials for this article are available online.

Comparing the Slack-Variable Mixture Model With Other Alternatives

P. 255-268

Lulu Kang, Javier Cruz Salgado & William A. Brenneman

Abstract

There have been many linear regression models proposed to analyze mixture experiments including the Scheffé model, the slack-variable model, and the Kronecker model. The use of the slack-variable model is somewhat controversial within the mixture experiment research community. However, in situations that the slack-variable ingredient is used to fill in the formulation and the remaining ingredients have constraints such that they can be chosen independently of one another, the slack-variable model is extremely popular by practitioners mainly due to the ease of interpretation. In this article, we advocate that for some mixture experiments the slack-variable model has appealing properties including numerical stability and better prediction accuracy when model-term selection is performed. We also explain how the effects of the slack-variable model components should be interpreted and how easy it is for practitioners to understand the components effects. We also investigate how to choose the slack-variable component, what transformation should be used to reduce collinearity, and under what circumstances the slack-variable model should be preferred. Both simulation and practical examples are provided to support the conclusions.

Optimal Experimental Designs in the Flow Rate of Particles

P. 269-276

Mariano Amo-Salas, Elvira Delgado-Márquez & Jesús López-Fidalgo

Abstract

This article focuses on analyzing the process of jam formation during the discharge by gravity of granular material stored in a two-dimensional silo. The aim of the article is two-fold. First, optimal experimental designs are computed, in which four approaches are considered: D-optimality, a combination of D-optimality and a cost/gain function, Bayesian D-optimality, and sequential designing. These results reveal that the efficiency of the design used by the experimenters can be improved dramatically. A sensitivity analysis with respect to the most important parameter is also performed. Second, estimation of the unknown parameters is done using least squares, that is, assuming normality, and also via maximum likelihood assuming the exponential distribution. Simulations for the designs considered in this article show that the variance, the mean squared error, and the bias of the estimators using maximum likelihood are in most cases lower than those using least squares. Supplementary materials for this article are available online.

Technometrics

A Journal of Statistics for the Physical, Chemical and Engineering Sciences

ASQ and American Statistical Association

Technometrics, ISSN 0040-1706
Volume 58, number 3 (august 2016): "Special Issue on Big Data"

Orthogonalizing EM: A Design-Based Least Squares Algorithm P. 285-293
Shifeng Xiong, Bin Dai, Jared Huling & Peter Z. G. Qian

Abstract

We introduce an efficient iterative algorithm, intended for various least squares problems, based on a design of experiments perspective. The algorithm, called orthogonalizing EM (OEM), works for ordinary least squares (OLS) and can be easily extended to penalized least squares. The main idea of the procedure is to orthogonalize a design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. We establish several attractive theoretical properties concerning OEM. For the OLS with a singular regression matrix, an OEM sequence converges to the Moore-Penrose generalized inverse-based least squares estimator. For ordinary and penalized least squares with various penalties, it converges to a point having grouping coherence for fully aliased regression matrices. Convergence and the convergence rate of the algorithm are examined. Finally, we demonstrate that OEM is highly efficient for large-scale least squares and penalized least squares problems, and is considerably faster than competing methods when n is much larger than p. Supplementary materials for this article are available online.

Speeding Up Neighborhood Search in Local Gaussian Process Prediction P. 294-303
Robert B. Gramacy & Benjamin Haaland

Abstract

Recent implementations of local approximate Gaussian process models have pushed computational boundaries for nonlinear, nonparametric prediction problems, particularly when deployed as emulators for computer experiments. Their flavor of spatially independent computation accommodates massive parallelization, meaning that they can handle designs two or more orders of magnitude larger than previously. However, accomplishing that feat can still require massive computational horsepower. Here we aim to ease that burden. We study how predictive variance is reduced as local designs are built up for prediction. We then observe how the exhaustive and discrete nature of an important search subroutine involved in building such local designs may be overly conservative. Rather, we suggest that searching the space radially, that is, continuously along rays emanating from the predictive location of interest, is a far thriftier alternative. Our empirical work demonstrates that ray-based search yields predictors with accuracy comparable to exhaustive search, but in a fraction of the time—for many problems bringing a supercomputer implementation back onto the desktop. Supplementary materials for this article are available online.

A Bootstrap Metropolis–Hastings Algorithm for Bayesian Analysis of Big Data P. 304-318
Faming Liang, Jinsu Kim & Qifan Song

Abstract

RMarkov chain Monte Carlo (MCMC) methods have proven to be a very powerful tool for analyzing data of complex structures. However, their computer-intensive nature, which typically require a large number of iterations and a complete scan of the full dataset for each iteration, precludes their use for big data analysis. In this article, we propose the so-called bootstrap Metropolis–Hastings (BMH) algorithm that provides a general framework for how to tame powerful MCMC methods to be used for big data analysis, that is, to replace the full data log-likelihood by a Monte Carlo average of the log-likelihoods that are calculated in parallel from multiple bootstrap samples. The BMH algorithm possesses an embarrassingly parallel structure and avoids repeated scans of the full dataset in iterations, and is thus feasible for big data problems. Compared to the popular divide-and-combine method, BMH can be generally more efficient as it can asymptotically integrate the whole data information into a single simulation run. The BMH algorithm is very flexible. Like the Metropolis–Hastings algorithm, it can serve as a basic building block for developing advanced MCMC algorithms that are feasible for big data problems. This is illustrated in the article by the tempering BMH algorithm, which can be viewed as a combination of parallel tempering and the BMH algorithm. BMH can also be used for model selection and optimization by combining with reversible jump MCMC and simulated annealing, respectively. Supplementary materials for this article are available online.

Compressing an Ensemble With Statistical Models: An Algorithm for Global 3D Spatio-Temporal Temperature

P. 319-328

Stefano Castruccio & Marc G. Genton

Abstract

One of the main challenges when working with modern climate model ensembles is the increasingly larger size of the data produced, and the consequent difficulty in storing large amounts of spatio-temporally resolved information. Many compression algorithms can be used to mitigate this problem, but since they are designed to compress generic scientific datasets, they do not account for the nature of climate model output and they compress only individual simulations. In this work, we propose a different, statistics-based approach that explicitly accounts for the space-time dependence of the data for annual global three-dimensional temperature fields in an initial condition ensemble. The set of estimated parameters is small (compared to the data size) and can be regarded as a summary of the essential structure of the ensemble output; therefore, it can be used to instantaneously reproduce the temperature fields in an ensemble with a substantial saving in storage and time. The statistical model exploits the gridded geometry of the data and parallelization across processors. It is therefore computationally convenient and allows to fit a nontrivial model to a dataset of 1 billion data points with a covariance matrix comprising of 1018entries. Supplementary materials for this article are available online.

Partitioning a Large Simulation as It Runs

P. 329-340

Kary Myers, Earl Lawrence, Michael Fugate, Claire McKay Bowen, Lawrence Ticknor, Jon Woodring, Joanne Wendelberger & Jim Ahrens

Abstract

As computer simulations continue to grow in size and complexity, they present a particularly challenging class of big data problems. Many application areas are moving toward *exascale* computing systems, systems that perform 1018 FLOPS (FLoating-point Operations Per Second)—a billion billion calculations per second. Simulations at this scale can generate output that exceeds both the storage capacity and the bandwidth available for transfer to storage, making post-processing and analysis challenging. One approach is to embed some analyses in the simulation while the simulation is running—a strategy often called *in situ analysis*—to reduce the need for transfer to storage. Another strategy is to save only a reduced set of time steps rather than the full simulation. Typically the selected time steps are evenly spaced, where the spacing can be defined by the budget for storage and transfer. This article combines these two ideas to introduce an online in situ method for identifying a reduced set of time steps of the simulation to save. Our approach significantly reduces the data transfer and storage requirements, and it provides improved fidelity to the simulation to facilitate post-processing and reconstruction. We illustrate the method using a computer simulation that supported NASA's 2009 Lunar Crater Observation and Sensing Satellite mission.

High-Performance Kernel Machines With Implicit Distributed Optimization and Randomization

P. 341-349

Haim Avron & Vikas Sindhwani

Abstract

We propose a framework for massive-scale training of kernel-based statistical models, based on combining distributed convex optimization with randomization techniques. Our approach is based on a block-splitting variant of the alternating directions method of multipliers, carefully reconfigured to handle very large random feature matrices under memory constraints, while exploiting hybrid parallelism typically found in modern clusters of multicore machines. Our high-performance implementation supports a variety of statistical learning tasks by enabling several loss functions, regularization schemes, kernels, and layers of randomized approximations for both dense and sparse datasets, in an extensible framework. We evaluate our implementation on large-scale model construction tasks and provide a comparison against existing sequential and parallel libraries. Supplementary materials for this article are available online.

Statistical Learning of Neuronal Functional Connectivity

P. 350-359

Chunming Zhang, Yi Chai, Xiao Guo, Muhong Gao, David Devilbiss & Zhengjun Zhang

Abstract

Identifying the network structure of a neuron ensemble beyond the standard measure of pairwise correlations is critical for understanding how information is transferred within such a neural population. However, the spike train data pose significant challenges to conventional statistical methods due to not only the complexity, massive size, and large scale, but also high dimensionality. In this article, we propose a novel "structural information enhanced" (SIE) regularization method for estimating the conditional intensities under the generalized linear model (GLM) framework to better capture the functional connectivity among neurons. We study the consistency of parameter estimation of the proposed method. A new "accelerated full gradient update" algorithm is developed to efficiently handle the complex penalty in the SIE-GLM for large sparse datasets applicable to spike train data. Simulation results indicate that our proposed method outperforms existing approaches. An application of the proposed method to a real spike train dataset, obtained from the prelimbic region of the prefrontal cortex of adult male rats when performing a T-maze based delayed-alternation task of working memory, provides some insight into the neuronal network in that region.

Measuring Influence of Users in Twitter Ecosystems Using a Counting Process Modeling Framework

P. 360-370

Donggeng Xia, Shawn Mankad & George Michailidis

Abstract

Data extracted from social media platforms are both large in scale and complex in nature, since they contain both unstructured text, as well as structured data, such as time stamps and interactions between users. A key question for such platforms is to determine influential users, in the sense that they generate interactions between members of the platform. Common measures used both in the academic literature and by companies that provide analytics services are variants of the popular web-search PageRank algorithm applied to networks that capture connections between users. In this work, we develop a modeling framework using multivariate interacting counting processes to capture the detailed actions that users undertake on such platforms, namely posting original content, reposting and/or mentioning other users' postings. Based on the proposed model, we also derive a novel influence measure. We discuss estimation of the model parameters through maximum likelihood and establish their asymptotic properties. The proposed model and the accompanying influence measure are illustrated on a dataset covering a five-year period of the Twitter actions of the members of the U.S. Senate, as well as mainstream news organizations and media personalities. Supplementary material is available online including computer code, data, and derivation details.

Discovering the Nature of Variation in Nonlinear Profile Data

P. 371-382

Zhenyu Shi, Daniel W. Apley & George C. Runger

Abstract

Profile data have received substantial attention in the quality control literature. Most of the prior work has focused on the profile monitoring problem of detecting sudden changes in the characteristics of the profiles, relative to an in-control sample set of profiles. In this article, we present an approach for exploratory analysis of a sample of profiles, the purpose of which is to discover the nature of any profile-to-profile variation that is present over the sample. This is especially challenging in big data environments in which the sample consists of a stream of high-dimensional profiles, such as image or point cloud data. We use manifold learning methods to find a low-dimensional representation of the variation, followed by a supervised learning step to map the low-dimensional representation back into the profile space. The mapping can be used for graphical animation and visualization of the nature of the variation, to facilitate root cause diagnosis. Although this mapping is related to a nonlinear mixed model sometimes used in profile monitoring, our focus is on discovering an appropriate characterization of the profile-to-profile variation, rather than assuming some prespecified parametric profile model and monitoring for variation in those specific parameters. We illustrate with two examples and include an additional example in the online supplement to this article on the *Technometrics* website.

Variable Selection in a Log–Linear Birnbaum–Saunders Regression Model for High-Dimensional Survival Data via the Elastic-Net and Stochastic EM

P. 383-392

Yukun Zhang, Xuewen Lu & Anthony F. Desmond

Abstract

The Birnbaum–Saunders (BS) distribution is broadly used to model failure times in reliability and survival analysis. In this article, we propose a simultaneous parameter estimation and variable selection procedure in a log–linear BS regression model for high-dimensional survival data. To deal with censored survival data, we iteratively run a combination of the stochastic EM algorithm (SEM) and variable selection procedure to generate pseudo-complete data and select variables until convergence. Treating pseudo-complete data as uncensored data via SEM makes it possible to incorporate iterative penalized least squares and simplify computation. We demonstrate the efficacy of our method using simulated and real datasets.

Online Updating of Statistical Inference in the Big Data Setting

P. 393-403

Elizabeth D. Schifano, Jing Wu, Chun Wang, Jun Yan & Ming-Hui Chen

Abstract

We present statistical methods for big data arising from online analytical processing, where large amounts of data arrive in streams and require fast analysis without storage/access to the historical data. In particular, we develop iterative estimating algorithms and statistical inferences for linear models and estimating equations that update as new data arrive. These algorithms are computationally efficient, minimally storage-intensive, and allow for possible rank deficiencies in the subset design matrices due to rare-event covariates. Within the linear model setting, the proposed online-updating framework leads to predictive residual tests that can be used to assess the goodness of fit of the hypothesized model. We also propose a new online-updating estimator under the estimating equation setting. Theoretical properties of the goodness-of-fit tests and proposed estimators are examined in detail. In simulation studies and real data applications, our estimator compares favorably with competing approaches under the estimating equation setting. Supplementary materials for this article are available online.