Biblioteca del Instituto de Estadística y Cartografía de Andalucía

Resúmenes de revistas
Julio - Septiembre 2023

Junta de Andalucía
Consejería de Economía, Hacienda y Fondos Europeos
Instituto de Estadística y Cartografía de Andalucía

PRESENTACIÓN

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

Los resúmenes de este boletín corresponden a las revistas que han ingresado en la Biblioteca del Instituto de Estadística y Cartografía de Andalucía durante el período **de julio a septiembre de 2023** 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:
Jueves: de 9:00h a 14:00h. y de 16:00 a 19:00 h.
Lunes, martes, miércoles y viernes: de 9:00h a 14:00h.
Horario de verano (del 15 de junio al 15 de septiembre), Semana Santa, Feria de Sevilla y Navidad (del 24 de diciembre al 6 de enero): de lunes a viernes de 9:00h. a 14:00h.

AH: Andalucía en la historia, ISSN 1695-1956
Número 77 (octubre – diciembre 2022)

Resumen
En 2022 se cumple el bicentenario de los gobiernos de dos liberales andaluces: Ramón Olaguer Feliú y, especialmente, Francisco Martínez de la Rosa, el hombre que haría posible el tránsito definitivo de la Monarquía absoluta a la Monarquía constitucional. Ambos lideraron una generación no solo de políticos y gestores sino de teóricos andaluces del gobierno representativo que replantearon por completo las bases doctrinales del liberalismo doceañista para hacer compatible la Corona con el régimen constitucional, en una línea bastante semejante al pensamiento posrevolucionario francés. Este monográfico, coordinado por Roberto Villa García, profesor titular de Historia Política en la Universidad Rey Juan Carlos, reúne las trayectorias biográficas y el pensamiento político de los liberales andaluces subrayando sus contribuciones al asentamiento de la Monarquía liberal en España

Resumen
Ante el desconocimiento de restos de los amurallamientos de la Sevilla anteriores al siglo XII, un nuevo hallazgo, el descubrimiento de un tramo amurallado del Bajo Imperio, fechado de finales del siglo III, tiene importantes consecuencias a la hora de abordar el estudio de la ciudad tardoantigua. Este reciente hallazgo arqueológico nos obliga, por tanto, a replantearnos el conocimiento del urbanismo hispalense.

Resumen
María Pacheco, esposa del comunero Juan de Padilla, fue un exponente claro del valor y la osadía que la llevó a enfrentarse no solo al emperador Carlos V, sino también a su propia familia. La historia se empeñó en dirigir la atención a su actitud política, pero no al latido precursor de un sentimiento feminista que estaba aún lejos de encontrar su tiempo para ser una militancia. Quinientos años después de su exilio en Portugal parece el momento idóneo de hacer una nueva lectura de su vida.

Resumen

Mirar un mapa antiguo nos pone el ojo en la historia. Refleja la visión del espacio geográfico y la capacidad técnica para construirlo en otros tiempos, representa la percepción que se tenía de las costas, relieves, ríos y otros accidentes, desentraña la organización del territorio y distribución del poblamiento, informa de la toponimia que se empleaba, e ilustra la riqueza artística de la cartografía antigua, denotando múltiples facetas políticas , sociales y culturales. Así sucede al contemplar el espléndido mapa de Andalucía del cartógrafo italiano Giacomo Cantelli publicado a finales del siglo XVII, en el que ciencia y arte corren de la mano

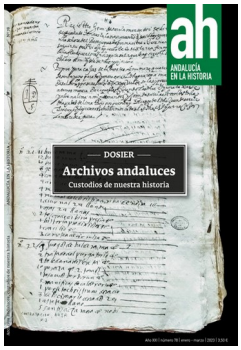
Resumen
Cuando en 1776 las 13 repúblicas rebeldes americanas del Reino Unido declaran la independencia de su metrópoli se inicia el camino del país que hoy conocemos como los Estados Unidos de América. Pocos años después el que luego será su presidente, Thomas Jefferson, uno de los que firma la Declaración de Independencia, nombrará al primer cónsul en España, concremente en Cádiz, a la par de Bilbao, en 1790. Y es que los "nuevos" norteamericanos se aprovecharon de su neutralidad en las guerras napoleónicas para continuar con su comercio atlántico y mediterráneo, usando el puerto de Cádiz como una de las escalas imprescindibles en la carga y descarga de mercadurías y personas. En este artículo hablamos también de otros vínculos con Cádiz, como el apoyo español al proceso de independencia de EE.UU.

Resumen
La relación que durante años mantuvieron la escritora Gertrudis Gómez de Avellaneda y el hacendado Ignacio de Cepeda salió a la luz en 1907, cuando se publicó la correspondencia secreta que habían iniciado en el verano de 1839. Superado el escándalo que causó entonces este descubrimiento y para admiración de la crítica literaria, las reediciones de estas cartas han superado en número a las de la extensa obra de la autora hispano-cubana, pese a que muy pronto se hizo evidente que su editor las había manipulado a conciencia para hacer de ella una heroína romántica.

Resumen
Más allá del tradicional periodismo político que conservará su predominio durante toda la Restauración, poco a poco, la prensa moderada fue incorporando nuevos temas y enfoques como estrategia para atraer a más lectores. Uno de ellos fueron las crónicas de sucesos que, a día de hoy, entendidas corno fuente histórica, nos permiten acceder a la vida cotidiana protagonizada por las clases sociales tradicionalmente silenciadas por la historiografía. En el caso de Málaga, fue el rotativo La Unión Mercantil (1886-1923), ejemplo de diario burgués de marcado carácter empresarial y el más importante de la ciudad en sus 50 años de existencia, el más destacado en la publicación de ese tipo de informaciones.

Resumen
Durante la Primera Guerra Mundial Alemania perdió todas sus colonias en África. En Camerún los ejércitos aliados arrinconaron a las tropas alemanas en la frontera con la vecina colonia de Guinea española. Miles de soldados y

civiles, tanto alemanes como cameruneses, huyeron del avance aliado y atravesaron la frontera del Río Muni en febrero de 1916. El gobierno español, neutral en la guerra, decidió acoger a estos refugiados de guerra. Los alemanes fueron trasladados a la Península. Entre ellos viajaron una serie de misioneros que se asentaron en la provincia de Cádiz. Libres de las disposiciones que afectaban al personal militar estos religiosos llamaron la atención de las autoridades españolas por sus actividades relacionadas con el espionaje en distintos pueblos y lugares de la provincia gaditana.



AH: Andalucía en la historia, ISSN 1695-1956
Número 78 (enero – marzo 2023)

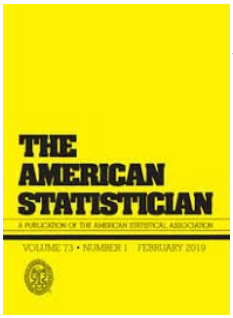
Resumen
En Andalucía existen más de 2.000 archivos públicos y privados, a los que hay que sumar los archivos de entidades religiosas, empresas, industrias y otros organismos y asociaciones. Abiertos a los investigadores y a los ciudadanos interesados en nuestro ayer, su trabajo es fundamental para el estudio de nuestro pasado y la preservación de nuestro presente para las generaciones venideras. Pionera en establecer normas para regular su acceso y conservación, merced a la aprobación de la Ley de Archivos en 1984, Andalucía también ha estado a la vanguardia de la digitalización de sus fondos con un doble objetivo: preservar y difundir nuestro legado. Conscientes de que sin su existencia y sin el trabajo silencioso de los archiveros la publicación de esta revista no sería posible, con ocasión de los 20 años de vida de Andalucía en la Historia en enero de 2023, les dedicamos este dosier coordinado por José Escalante, director del Archivo Municipal de Antequera, y Mercedes Fernández Paradas, profesora titular de Historia Contemporánea de la Universidad de Málaga.

Resumen
En los centros comerciales de las grandes ciudades del siglo XXI no faltan tiendas de óptica. Cuesta imaginar, pero hay que hacerlo, que en una Sevilla o Sanlúcar del año 1519, el maestre Juan Sebastián Elcano antes de tomar rumbo a las Molucas, necesitó graduarse la vista y comprar unos antojos en las alcaicerías de estas bulliciosas ciudades andaluzas. Grabados y cuadros de la época nos permiten conocer cómo eran esos comercios.

Resumen
El libro Magia natural, o ciencia de filosofía oculta, con nuevas noticias de los más profundos misterios y secretos del universo visible, obra del jesuíta gaditano Hernando Castrillo (1585-1667), salido en 1649 de la imprenta de Diego Pérez Estupiñán, fue el primer libro impreso en tierras de Huelva, concretamente en la villa de Trigueros. Hubo que esperar nada menos que 180 años, para que viese la luz el segundo libro impreso en la provincia.

Resumen
Este artículo relata una historia apasionante. La aventura de dos obreros andaluces de la Rio Tinto Company que fueron reclutados en 1910 para trabajar como "expertos extranjeros" en las fundiciones de cobre que el gobierno chino estaba abriendo en la provincia de Gansu, en la última frontera China. Un lugar tan alejado, que es donde hoy China hace sus pruebas nucleares y lanza sus cohetes al espacio. Pudieron pasar de colonizados a colonizadores, pero la suerte no les sonrió. La revolución de 1911 les obligó a huir, recorriendo los abandonados senderos de la Ruta de la Seda hasta alcanzar el Transiberiano.

Resumen
En el pasdo mes de julio se cumplieron 50 años de la muerte del gran historiador y filólogo granadino Américo Castro. Exiliado en Estados Unidos desde 1938, protagonizó, junto al historiador español exiliado en Argentina, Claudio Sánchez Albornoz, una de las más conocidas polémicas historiográficasde nuestro pasado. En este artículo seguimos las huellas andaluzas de quien fuera alumno del rondeño Giner de los Ríos: Granada, Ángel Ganivet, la Alhambra, Luis Rosales, Emilio Orozco, Antonio Domínguez Ortiz, Antonio Gallego Morel, etc.



The American Statistician, ISSN 0003-1305
Volume 77, number 2 (may 2023)

Abstract
Reproducibility, the ability to reproduce the results of published papers or studies using their computer code and data, is a cornerstone of reliable scientific methodology. Studies where results cannot be reproduced by the scientific community should be treated with caution. Over the past decade, the importance of reproducible research has been frequently stressed in a wide range of scientific journals such as *Nature* and *Science* and international magazines such as *The Economist*. However, multiple studies have demonstrated that scientific results are often not reproducible across research areas such as psychology and medicine. Statistics, the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data, prides itself on its openness when it comes to sharing both computer code and data. In this article, we examine reproducibility in the field of statistics by attempting to reproduce the results in 93 published papers in prominent journals using functional magnetic resonance imaging (fMRI) data during the 2010–2021 period. Overall, from both the computer code and the data perspective, among all the 93 examined papers, we could only reproduce the results in 14 (15.1%) papers, that is, the papers provide both executable computer code (or software) with the real fMRI data, and our results matched the results in the paper. Finally, we conclude with some author-specific and journal-specific recommendations to improve the research reproducibility in statistics.

Abstract
We address the question of how to perform a paired *t*-test in situations where we do not know how to pair the data. Specifically, we discuss approaches for bounding the test statistic of the paired *t*-test in a way that allows us to recover the results of this test in some cases. We also discuss the relationship between the paired *t*-test and the independent samples *t*-test and what happens if we use the latter to approximate the former. Our results are informed by both theoretical results and a simulation study.

Abstract
Combining individual *p*-values to perform an overall test is often encountered in statistical applications. The Cauchy combination test (CCT) (*Journal of the American Statistical Association*, 2020, 115, 393–402) is a powerful and computationally efficient approach to integrate individual *p*-values under arbitrary dependence structures for sparse signals. We revisit this test to additionally show that (i) the tail probability of the CCT can be approximated just as well when more relaxed assumptions are imposed on individual *p*-values compared to those of the original test statistics; (ii) such assumptions are satisfied by six popular copula distributions; and (iii) the power of the CCT is no less than

that of the minimum *p*-value test when the number of *p*-values goes to infinity under some regularity conditions. These findings are confirmed by both simulations and applications in two real datasets, thus, further broadening the theory and applications of the CCT.

Bayesian-Frequentist Hybrid Inference in Applications with Small Sample Sizes P. 143-150

Gang Han, Thomas J. Santner, Haiqun Lin & Ao Yuan

Abstract

The Bayesian-frequentist hybrid model and associated inference can combine the advantages of both Bayesian and frequentist methods and avoid their limitations. However, except for few special cases in existing literature, the computation under the hybrid model is generally nontrivial or even unsolvable. This article develops a computation algorithm for hybrid inference under any general loss functions. Three simulation examples demonstrate that hybrid inference can improve upon frequentist inference by incorporating valuable prior information, and also improve Bayesian inference based on non-informative priors where the latter leads to biased estimates for the small sample sizes used in inference. The proposed method is illustrated in applications including a biomechanical engineering design and a surgical treatment of acral lentiginous melanoma.

Optimal and Fast Confidence Intervals for Hypergeometric Successes P. 151-159

Jay Bartroff, Gary Lorden & Lijia Wang

Abstract

We present an efficient method of calculating exact confidence intervals for the hypergeometric parameter representing the number of "successes," or "special items," in the population. The method inverts minimum-width acceptance intervals after shifting them to make their endpoints nondecreasing while preserving their level. The resulting set of confidence intervals achieves minimum possible average size, and even in comparison with confidence sets not required to be intervals it attains the minimum possible cardinality most of the time, and always within 1. The method compares favorably with existing methods not only in the size of the intervals but also in the time required to compute them. The available R package hyperMCI implements the proposed method.

Forbidden Knowledge and Specialized Training: A Versatile Solution for the Two Main Sources of Overfitting in Linear Regression P. 160-168

Chris Rohlfs

Abstract

Overfitting in linear regression is broken down into two main causes. First, the formula for the estimator includes "forbidden knowledge" about training observations' residuals, and it loses this advantage when deployed out-of-sample. Second, the estimator has "specialized training" that makes it particularly capable of explaining movements in the predictors that are idiosyncratic to the training sample. An out-of-sample counterpart is introduced to the popular "leverage" measure of training observations' importance. A new method is proposed to forecast out-of-sample fit at the time of deployment, when the values for the predictors are known but the true outcome variable is not. In Monte Carlo simulations and in an empirical application using MRI brain scans, the proposed estimator performs comparably to Predicted Residual Error Sum of Squares (PRESS) for the average out-of-sample case and unlike PRESS, also performs consistently across different test samples, even those that differ substantially from the training set.

Estimating Knee Movement Patterns of Recreational Runners Across Training Sessions Using Multilevel Functional Regression Models P. 169-181

Marcos Matabuena, Marta Karas, Sherveen Riazati, Nick Caplan & Philip R. Hayes

Abstract

Modern wearable monitors and laboratory equipment allow the recording of high-frequency data that can be used to quantify human movement. However, currently, data analysis approaches in these domains remain limited. This article proposes a new framework to analyze biomechanical patterns in sport training data recorded across multiple training sessions using multilevel functional models. We apply the methods to subsecond-level data of knee location trajectories collected in 19 recreational runners during a medium-intensity continuous run (MICR) and a high-intensity interval training (HIIT) session, with multiple steps recorded in each participant-session. We estimate functional intra-class correlation coefficient to evaluate the reliability of recorded measurements across multiple sessions of the same training type. Furthermore, we obtained a vectorial representation of the three hierarchical levels of the data and visualize them in a low-dimensional space. Finally, we quantified the differences between genders and between two training types using functional multilevel regression models that incorporate covariate information. We provide an overview of the relevant methods and make both data and the R code for all analyses freely available online on GitHub. Thus, this work can serve as a helpful reference for practitioners and guide for a broader audience of researchers interested in modeling repeated functional measures at different resolution levels in the context of biomechanics and sports science applications.

Abstract

Conventional wisdom dispersed by fans and coaches in the stands at almost any high school track meet suggests female athletes typically peak around 10th grade or earlier (15 years of age), particularly for distance runners, and male athletes continuously improve. Given that universities in the United States typically recruit track and field athletes from high school teams, it is important to understand the age of peak performance at the high school level. Athletes are often recruited starting in their sophomore year of high school and individuals develop at different rates during adolescence; however, the individual development factor is usually not taken into account during recruitment. In this study, we curate data on event times for high school track and field athletes from the years 2011 to 2019 to determine the trajectory of fastest times for male and female athletes in the 200m, 400m, 800m, and 1600m races. We show, through visualizations and models, that, for most athletes, the sophomore peak is a myth. Performance is mostly dependent on the individual athlete. That said, the trajectories cluster into four or five types, depending on the race distance. We explain the significance of the types for future recruitment.

Abstract

When releasing record-level data containing sensitive information to the public, the data disseminator is responsible for protecting the privacy of every record in the dataset, simultaneously preserving important features of the data for users' analyses. These goals can be achieved by data synthesis, where confidential data are replaced with synthetic data that are simulated based on statistical models estimated on the confidential data. In this article, we present a data synthesis case study, where synthetic values of price and the number of available days in a sample of the New York Airbnb Open Data are created for privacy protection. One sensitive variable, the number of available days of an Airbnb listing, has a large amount of zero-valued records and also truncated at the two ends. We propose a zero-inflated truncated Poisson regression model for its synthesis. We use a sequential synthesis approach to further synthesize the sensitive price variable. The resulting synthetic data are evaluated for its utility preservation and privacy protection, the latter in the form of disclosure risks. Furthermore, we propose methods to investigate how uncertainties in intruder's knowledge would influence the identification disclosure risks of the synthetic data. In particular, we explore several realistic scenarios of uncertainties in intruder's knowledge of available information and evaluate their impacts on the resulting identification disclosure risks.

Abstract
Hierarchical clustering is one of the standard methods taught for identifying and exploring the underlying structures that may be present within a dataset. Students are shown examples in which the dendrogram, a visual representation of the hierarchical clustering, reveals a clear clustering structure. However, in practice, data analysts today frequently encounter datasets whose large scale undermines the usefulness of the dendrogram as a visualization tool. Densely packed branches obscure structure, and overlapping labels are impossible to read. In this article we present a new workflow for performing hierarchical clustering via the R package called protoshiny that aims to restore hierarchical clustering to its former role of being an effective and versatile visualization tool. Our proposal leverages interactivity combined with the ability to label internal nodes in a dendrogram with a representative data point (called a *prototype*). After presenting the workflow, we provide three case studies to demonstrate its utility.

Abstract
We provide a case study for motivating and teaching nonparametric statistical inference alongside traditional parametric approaches. The case consists of analyses by Bracht et al. who use analysis of variance (ANOVA) to assess the applicability of the human microfibrillar-associated protein 4 (MFAP4) as a biomarker for hepatic fibrosis in hepatitis C patients. We revisit their analyses and consider two nonparametric approaches: Mood's median test and the Kruskal-Wallis test. We demonstrate how this case study enables instructors to discuss critical assumptions of parametric procedures while comparing and contrasting the results of multiple approaches. Interestingly, only one of the three approaches creates groupings that match the treatment recommendations of the European Association for the Study of the Liver (EASL). We provide guidance and resources to aid instructors in directing their students through this case study at various levels, including R code and novel R shiny applications for conducting the analyses in the classroom.

Abstract
We recognize the careful reading of and thought-provoking commentary on our work by Rice and Lumley. Further, we appreciate the opportunity to respond and clarify our position regarding the three presented concerns. We address these points in three sections below and conclude with final remarks in Section 4.

Abstract
Reinforcement learning (RL) is a computational approach to reward-driven learning in sequential decision problems. It implements the discovery of optimal actions by learning from an agent interacting with an environment rather than from supervised data. We contrast and compare RL with traditional sequential design, focusing on simulation-based Bayesian sequential design (BSD). Recently, there has been an increasing interest in RL techniques for healthcare applications. We introduce two related applications as motivating examples. In both applications, the sequential nature of the decisions is restricted to sequential stopping. Rather than a comprehensive survey, the focus of the discussion is on solutions using standard tools for these two relatively simple sequential stopping problems. Both problems are inspired by adaptive clinical trial design. We use examples to explain the terminology and mathematical

background that underlie each framework and map one to the other. The implementations and results illustrate the many similarities between RL and BSD. The results motivate the discussion of the potential strengths and limitations of each approach.

Technometrics, ISSN 0040-1706
Volume 64, number 2 (may 2022)

Abstract
Classification is a major tool of statistics and machine learning. A classification method first processes a training set of objects with given classes (labels), with the goal of afterward assigning new objects to one of these classes. When running the resulting prediction method on the training data or on test data, it can happen that an object is predicted to lie in a class that differs from its given label. This is sometimes called label bias, and raises the question whether the object was mislabeled. The proposed class map reflects the probability that an object belongs to an alternative class, how far it is from the other objects in its given class, and whether some objects lie far from all classes. The goal is to visualize aspects of the classification results to obtain insight in the data. The display is constructed for discriminant analysis, the k-nearest neighbor classifier, support vector machines, logistic regression, and coupling pairwise classifications. It is illustrated on several benchmark datasets, including some about images and texts.

Abstract
In this article, we propose an optimal method referred to as SPlit for splitting a dataset into training and testing sets. SPlit is based on the method of support points (SP), which was initially developed for finding the optimal representative points of a continuous distribution. We adapt SP for subsampling from a dataset using a sequential nearest neighbor algorithm. We also extend SP to deal with categorical variables so that SPlit can be applied to both regression and classification problems. The implementation of SPlit on real datasets shows substantial improvement in the worst-case testing performance for several modeling methods compared to the commonly used random splitting procedure.

Abstract
Motivated by the wind turbine anomaly detection, we propose a Bayesian hierarchical model (BHM) for the mean-change detection in multivariate sequences. By combining the exchange random order distribution induced from the Poisson–Dirichlet process and nonlocal priors, BHM exhibits satisfactory performance for mean-shift detection with multivariate sequences under different error distributions. In particular, BHM yields the smallest detection error compared with other competitive methods considered in the article. We use a local scan procedure to accelerate the computation, while the anomaly locations are determined by maximizing the posterior probability through dynamic programming. We establish consistency of the estimated number and locations of the change points and conduct extensive simulations to evaluate the BHM approach. Among the popular change point detection algorithms, BHM

yields the best performance for most of the datasets in terms of the F1 score for the wind turbine anomaly detection.

PICAR: An Efficient Extendable Approach for Fitting Hierarchical Spatial Models P. 187-198

Ben Seiyon Lee, Murali Haran

Abstract

Hierarchical spatial models are very flexible and popular for a vast array of applications in areas such as ecology, social science, public health, and atmospheric science. It is common to carry out Bayesian inference for these models via Markov chain Monte Carlo (MCMC). Each iteration of the MCMC algorithm is computationally expensive due to costly matrix operations. In addition, the MCMC algorithm needs to be run for more iterations because the strong cross-correlations among the spatial latent variables result in slow mixing Markov chains. To address these computational challenges, we propose a projection-based intrinsic conditional autoregression (PICAR) approach, which is a discretized and dimension-reduced representation of the underlying spatial random field using empirical basis functions on a triangular mesh. Our approach exhibits fast mixing as well as a considerable reduction in computational cost per iteration. PICAR is computationally efficient and scales well to high dimensions. It is also automated and easy to implement for a wide array of user-specified hierarchical spatial models. We show, via simulation studies, that our approach performs well in terms of parameter inference and prediction. We provide several examples to illustrate the applicability of our method, including (i) a high-dimensional cloud cover dataset that showcases its computational efficiency, (ii) a spatially varying coefficient model that demonstrates the ease of implementation of PICAR in the probabilistic programming languages stan and nimble, and (iii) a watershed survey example that illustrates how PICAR applies to models that are not amenable to efficient inference via existing methods.

Sequential Design of Multi-Fidelity Computer Experiments: Maximizing the Rate of Stepwise Uncertainty Reduction P. 199-209

Rémi Stroh, Julien Bect, Séverine Demeyer, Nicolas Fischer, Damien Marquis, Emmanuel Vazquez

Abstract

This article deals with the sequential design of experiments for (deterministic or stochastic) multi-fidelity numerical simulators, that is, simulators that offer control over the accuracy of simulation of the physical phenomenon or system under study. Accurate simulations usually entail a high computational effort, while coarse simulations are obtained at a lower cost. The cost can be measured, for example, by the run time of the simulator or the financial cost of the computing resources. In this setting, simulation results obtained at several levels of fidelity can be combined in order to estimate quantities of interest (the optimal value of the output, the probability that the output exceeds a given threshold, etc.) in an efficient manner. We propose a new Bayesian sequential strategy called maximal rate of stepwise uncertainty reduction (MR-SUR), that selects additional simulations to be performed by maximizing the ratio between the expected reduction of uncertainty and the cost of simulation. This generic strategy unifies several existing methods, and provides a principled approach to develop new ones. We assess its performance on several examples, including a computationally intensive problem of fire safety analysis where the quantity of interest is the probability of exceeding a tenability threshold during a building fire.

Monitoring Heterogeneous Multivariate Profiles Based on Heterogeneous Graphical Model P. 210-223

Hui Wu, Chen Zhang, Yan-Fu Li

Abstract

Process monitoring using profile data remains an important and challenging problem in various manufacturing industries. Motivated by an application case of motherboard testing processes, we develop a novel modeling and monitoring framework for heterogeneous multivariate profiles. In this framework, a heterogeneous graphical model is

constructed to depict the complicated heterogeneous relationship among profile channels. Then monitoring the heterogeneous relationship among profile channels can be reduced to monitoring the graphical networks. Besides, we investigate several theoretical results concerning the accuracy of the estimated graphical structure. Finally, we demonstrate the proposed method through extensive simulations and a real case study.

Bayesian Dynamic Feature Partitioning in High-Dimensional Regression With Big Data P. 224-240

Rene Gutierrez & Rajarshi Guhaniyogi

Abstract

Bayesian computation of high-dimensional linear regression models using Markov chain Monte Carlo (MCMC) or its variants can be extremely slow or completely prohibitive since these methods perform costly computations at each iteration of the sampling chain. Furthermore, this computational cost cannot usually be efficiently divided across a parallel architecture. These problems are aggravated if the data size is large or data arrive sequentially over time (streaming or online settings). This article proposes a novel dynamic feature partitioned regression (DFP) for efficient online inference for high-dimensional linear regressions with large or streaming data. DFP constructs a *pseudo posterior density* of the parameters at every time point, and quickly updates the pseudo posterior when a new block of data (data shard) arrives. DFP updates the pseudo posterior at every time point suitably and partitions the set of parameters to exploit parallelization for efficient posterior computation. The proposed approach is applied to high-dimensional linear regression models with Gaussian scale mixture priors and spike-and-slab priors on large parameter spaces, along with large data, and is found to yield state-of-the-art inferential performance. The algorithm enjoys theoretical support with pseudoposterior densities over time being arbitrarily close to the full posterior as the data size grows, as shown in the supplementary material. Supplementary material also contains details of the DFP algorithm applied to different priors. Package to implement DFP is available in https://github.com/Rene-Gutierrez/DynParRegReg. The dataset is available in https://github.com/Rene-Gutierrez/DynParRegReg_Implementation.

Anomaly Detection in Large-Scale Networks With Latent Space Models P. 241-252

Wesley Lee, Tyler H. McCormick, Joshua Neil, Cole Sodja, Yanran Cui

Abstract

We develop a real-time anomaly detection method for directed activity on large, sparse networks. We model the propensity for future activity using a dynamic logistic model with interaction terms for sender- and receiver-specific latent factors in addition to sender- and receiver-specific popularity scores; deviations from this underlying model constitute potential anomalies. Latent nodal attributes are estimated via a variational Bayesian approach and may change over time, representing natural shifts in network activity. Estimation is augmented with a case-control approximation to take advantage of the sparsity of the network and reduces computational complexity from $O(N2)$ to $O(E)$, where N is the number of nodes and E is the number of observed edges. We run our algorithm on network event records collected from an enterprise network of over 25,000 computers and are able to identify a red team attack with half the detection rate required of the model without latent interaction terms.

An Adaptive Sampling Strategy for Online Monitoring and Diagnosis of High-Dimensional Streaming Data P. 253-269

Ana María Estrada Gómez, Dan Li, Kamran Paynabar

Abstract

Statistical process control techniques have been widely used for online process monitoring and diagnosis of streaming data in various applications, including manufacturing, healthcare, and environmental engineering. In some applications, the sensing system that collects online data can only provide partial information from the process due to resource constraints. In such cases, an adaptive sampling strategy is needed to decide where to collect data while

maximizing the change detection capability. This article proposes an adaptive sampling strategy for online monitoring and diagnosis with partially observed data. The proposed methodology integrates two novel ideas (i) the recursive projection of the high-dimensional streaming data onto a low-dimensional subspace to capture the spatio-temporal structure of the data while performing missing data imputation; and (ii) the development of an adaptive sampling scheme, balancing exploration and exploitation, to decide where to collect data at each acquisition time. Through simulations and two case studies, the proposed framework's performance is evaluated and compared with benchmark methods.

Technometrics, ISSN 0040-1706
Volume 64, number 2 (may 2022)

Technometrics, ISSN 0040-1706
Volume 64, number 3 (august 2022)

Abstract
In this work, we predict the outcomes of high fidelity multivariate computer simulations from low fidelity counterparts using function-to-function regression. The high fidelity simulation takes place on a high definition mesh, while its low fidelity counterpart takes place on a coarsened and truncated mesh. We showcase our approach by applying it to a complex finite element simulation of an abdominal aortic aneurysm which provides the displacement field of a blood vessel under pressure. In order to link the two multidimensional outcomes we compress them and then fit a function-to-function regression model. The data are high dimensional but of low sample size, meaning that only a few simulations are available, while the output of both low and high fidelity simulations is in the order of several thousands. To match this specific condition our compression method assumes a Gaussian Markov random field that takes the finite element geometry into account and only needs little data. In order to solve the function-to-function regression model we construct an appropriate prior with a shrinkage parameter which follows naturally from a Bayesian view of the Karhunen–Loève decomposition. Our model enables real multivariate predictions on the complete grid instead of resorting to the outcome of specific points.

Abstract
Motivated by the Golub–Heath–Wahba formula for ridge regression, we first present a new leave-one-out lemma for the kernel support vector machines (SVM) and related large-margin classifiers. We then use the lemma to design a novel and efficient algorithm, named "magicsvm," for training the kernel SVM and related large-margin classifiers and computing the exact leave-one-out cross-validation error. By "magicsvm," the computational cost of leave-one-out analysis is of the same order of fitting a single SVM on the training data. We show that "magicsvm" is much faster than the state-of-the-art SVM solvers based on extensive simulations and benchmark examples. The same idea is also used to boost the computation speed of the *V*-fold cross-validation of the kernel classifiers.

Abstract
Bayesian calibration of a functional input/parameter to a time-consuming simulator based on a Gaussian process (GP) emulator involves two challenges that distinguish it from other parameter calibration problems. First, one needs to specify a flexible stochastic process prior for the input, and reduce it to a tractable number of random variables. Second, a sequential experiment design criterion that decreases the effect of emulator prediction uncertainty on

calibration results is needed and the criterion should be scalable for high-dimensional input and output. In this article, we address these two issues. For the first issue, we employ a GP with a prior density for its correlation parameter as prior for the functional input, and the Karhunen-Loève (KL) expansion of this non-Gaussian stochastic process to reduce its dimension. We show that this prior gives far more robust inference results than a GP with a fixed correlation parameter. For the second issue, we propose the weighted prediction variance (WPV) criterion (with posterior density of the calibration parameter as weight) and prove the consistency of the sequence of emulator-based likelihoods given by the criterion. The proposed method is illustrated with examples on hydraulic transmissivity estimation for groundwater models.

Data-Driven Determination of the Number of Jumps in Regression Curves P. 312-322

Guanghui Wang, Changliang Zou & Peihua Qiu

Abstract

In nonparametric regression with jump discontinuities, one major challenge is to determine the number of jumps in a regression curve. Most existing methods to solve that problem are based on either a sequence of hypothesis tests or model selection, by introducing some extra tuning parameters that may not be easy to determine in practice. This article aims to develop a data-driven new methodology for determining the number of jumps, using an order-preserved sample-splitting strategy together with a cross-validation-based criterion. Statistical consistency of the determined number of jumps by our proposed method is established. More interestingly, the proposed method allows us to move beyond just point estimation, and it can quantify uncertainty of the proposed estimate. The key idea behind our method is the construction of a series of statistics with marginal symmetry property and this property can be used for choosing a data-driven threshold to control the false discovery rate of our method. The proposed method is computationally efficient. Numerical experiments indicate that it has a reliable performance in finite-sample cases. An R package jra is developed to implement the proposed method.

Reliable Post-Signal Fault Diagnosis for Correlated High-Dimensional Data Streams P. 323-334

Dongdong Xiang, Peihua Qiu, Dezhi Wang & Wendong Li

Abstract

Rapid advance of sensor technology is facilitating the collection of high-dimensional data streams (HDS). Apart from real-time detection of potential out-of-control (OC) patterns, post-signal fault diagnosis of HDS is becoming increasingly important in the filed of statistical process control to isolate abnormal data streams. The major limitations of the existing methods on that topic include (i) they cannot achieve reliable diagnostic results in the sense that their performance is highly variable, and (ii) the informative correlation among different streams is often neglected by them. This article elaborates the problem of reliable fault diagnosis for monitoring correlated HDS using the large-scale multiple testing. Under the framework of hidden Markov model dependence, new diagnostic procedures are proposed, which can control the missed discovery exceedance (MDX) at a desired level. Extensive numerical studies along with some theoretical results show that the proposed procedures can control MDX properly, leading to diagnostics with high reliability and efficiency. Also, their diagnostic performance can be improved significantly by exploiting the dependence among different data streams, which is especially appealing in practice for identifying clustered OC streams.

Functional PCA With Covariate-Dependent Mean and Covariance Structure P. 335-345

Fei Ding, Shiyuan He, David E. Jones & Jianhua Z. Huang

Abstract

Incorporating covariates into functional principal component analysis (PCA) can substantially improve the representation efficiency of the principal components and predictive performance. However, many existing functional

PCA methods do not make use of covariates, and those that do often have high computational cost or make overly simplistic assumptions that are violated in practice. In this article, we propose a new framework, called covariate-dependent functional principal component analysis (CD-FPCA), in which both the mean and covariance structure depend on covariates. We propose a corresponding estimation algorithm, which makes use of spline basis representations and roughness penalties, and is substantially more computationally efficient than competing approaches of adequate estimation and prediction accuracy. A key aspect of our work is our novel approach for modeling the covariance function and ensuring that it is symmetric positive semidefinite. We demonstrate the advantages of our methodology through a simulation study and an astronomical data analysis.

Spectral Clustering on Spherical Coordinates Under the Degree-Corrected Stochastic Blockmodel P. 346-357

Francesco Sanna Passino, Nicholas A. Heard & Patrick Rubin-Delanchy

Abstract

Spectral clustering is a popular method for community detection in network graphs: starting from a matrix representation of the graph, the nodes are clustered on a low-dimensional projection obtained from a truncated spectral decomposition of the matrix. Estimating correctly the number of communities and the dimension of the reduced latent space is critical for good performance of spectral clustering algorithms. Furthermore, many real-world graphs, such as enterprise computer networks studied in cyber-security applications, often display heterogeneous within-community degree distributions. Such heterogeneous degree distributions are usually not well captured by standard spectral clustering algorithms. In this article, a novel spectral clustering algorithm is proposed for community detection under the degree-corrected stochastic blockmodel. The proposed method is based on a transformation of the spectral embedding to spherical coordinates, and a novel modeling assumption in the transformed space. The method allows for simultaneous and automated selection of the number of communities and the latent dimension for spectral embeddings of graphs with uneven node degrees. Results show improved performance over competing methods in representing computer networks.

Locally Optimal Design for A/B Tests in the Presence of Covariates and Network Dependence P. 358-369

Qiong Zhang & Lulu Kang

Abstract

A/B test, a simple type of controlled experiment, refers to the statistical procedure of experimenting to compare two treatments applied to test subjects. For example, many IT companies frequently conduct A/B tests on their users who are connected and form social networks. Often, the users' responses could be related to the network connection. In this article, we assume that the users, or the test subjects of the experiments, are connected on an undirected network, and the responses of two connected users are correlated. We include the treatment assignment, covariate features, and network connection in a conditional autoregressive model. Based on this model, we propose a design criterion that measures the variance of the estimated treatment effect and allocate the treatment settings to the test subjects by minimizing the criterion. Since the design criterion depends on an unknown network correlation parameter, we adopt the locally optimal design method and develop a hybrid optimization approach to obtain the optimal design. Through synthetic and real social network examples, we demonstrate the value of including network dependence in designing A/B experiments and validate that the proposed locally optimal design is robust to the choices of parameters. Supplementary materials for this article are available online.

A Statistical Approach to Surface Metrology for 3D-Printed Stainless Steel P. 370-383

Chris J. Oates, Wilfrid S. Kendall & Liam Fleming

Abstract

The improvement of sensing technology enables features of process variables to be collected during the fabrication of

products. This article develops an automatic tool for process feature rankings based on these data. Based on the sensing data characteristics and the need of manufacturing system analysis, we propose two rules of the feature ranking scheme: assessing general dependency between each individual process feature and the quality variable, and satisfying a diversity rule. Specifically, we propose a feature ranking scheme based on the sparse distance correlation (SpaDC) that satisfies these two rules. Theoretical properties of the proposed algorithm are investigated. Simulation studies and two real-case studies from semiconductor manufacturing applications demonstrate that the SpaDC method ranks the features effectively given these two ranking rules.

Ranking Features to Promote Diversity: An Approach Based on Sparse Distance Correlation P. 384-395

Andi Wang, Juan Du, Xi Zhang & Jianjun Shi

Abstract

The improvement of sensing technology enables features of process variables to be collected during the fabrication of products. This article develops an automatic tool for process feature rankings based on these data. Based on the sensing data characteristics and the need of manufacturing system analysis, we propose two rules of the feature ranking scheme: assessing general dependency between each individual process feature and the quality variable, and satisfying a diversity rule. Specifically, we propose a feature ranking scheme based on the sparse distance correlation (SpaDC) that satisfies these two rules. Theoretical properties of the proposed algorithm are investigated. Simulation studies and two real-case studies from semiconductor manufacturing applications demonstrate that the SpaDC method ranks the features effectively given these two ranking rules.

Density Regression with Conditional Support Points P. 396-408

Yunlu Chen & Nan Zhang

Abstract

Density regression characterizes the conditional density of the response variable given the covariates, and provides much more information than the commonly used conditional mean or quantile regression. However, it is often computationally prohibitive in applications with massive datasets, especially when there are multiple covariates. In this article, we develop a new data reduction approach for the density regression problem using conditional support points. After obtaining the representative data, we exploit the penalized likelihood method as the downstream estimation strategy. Based on the connections among the continuous ranked probability score, the energy distance, the L_2 discrepancy and the symmetrized Kullback–Leibler distance, we investigate the distributional convergence of the representative points and establish the rate of convergence of the density regression estimator. The usefulness of the methodology is illustrated by modeling the conditional distribution of power output given multivariate environmental factors using a large scale wind turbine dataset.

A New Sparse-Learning Model for Maximum Gap Reduction of Composite Fuselage Assembly P. 409-418

Juan Du, Shanshan Cao, Jeffrey H. Hunt, Xiaoming Huo & Jianjun Shi

Abstract

Natural dimensional variabilities of incoming fuselages affect the assembly speed and quality of fuselage joins in composite fuselage assembly processes. Shape control is critical to ensure the quality of composite fuselage assembly. In current practice, the structures are adjusted to the design shape in terms of the l2 loss for further assembly without considering the existing dimensional gap between two structures. Such practice has two limitations: (a) controlling each fuselage to the design shape may not be the optimal shape control strategy in terms of a pair of incoming fuselages with different incoming dimensions; (b) the maximum gap is the key concern during the fuselage assembly process, so the l∞ loss of gap after control ought to be considered. This article proposes an optimal shape

control methodology via the l∞ loss for the composite fuselage assembly process by considering the existing dimensional gap between the incoming pair of fuselages. On the other hand, due to the limitation on the number of available actuators in practice, we face an important problem of finding the best locations for the actuators among many potential locations, which makes the problem a sparse estimation problem. We are the first to solve the optimal shape control in the fuselage assembly process using the l∞ model under the framework of sparse estimation, where we use the l1 penalty to control the sparsity of the resulting estimator. From the statistical point of view, this can be formulated as the l∞ loss based linear regression, and under some standard assumptions, such as the restricted eigenvalue (RE) conditions, and the light-tailed noise, the nonasymptotic estimation error of the l1 regularized l∞ linear model is derived, which meets the upper-bound in the existing literature. Compared to the current practice, the case study shows that our proposed method significantly reduces the maximum gap between two fuselages after shape adjustments.