Biblioteca del Instituto de Estadística y Cartografía de Andalucía

Resúmenes de revistas
Enero – Marzo 2025

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 **enero a marzo de 2025** 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:
Martes: de 9:00h a 14:00h. y de 16:00 a 19:00 h.
Lunes, miércoles, jueves y viernes: de 9:00h a 14:00h.
Horario de verano (del 15 de junio al 15 de septiembre), Semana Santa, Feria de Sevilla y Navidad (del 24 de diciembre al 6 de enero): de lunes a viernes de 9:00h. a 14:00h.

Cartographic journal, The, ISSN 0008-7041
Volume 61, number 1 (february 2024)

Abstract
Pieter Pourbus was a mid-sixteenth-century artist and mapmaker who crafted large-scale chorographic maps in the coastal area around Bruges (Flanders, The Low Countries). The topographic and planimetric accuracy of his maps has made researchers hypothesize that he applied the triangulation methods as first described by Gemma Frisius (1533). The actual application of triangulation for mapping purposes in the sixteenth century is however contested. A new macroscopic and GIS-integrated study of Pourbus' maps has revealed traces of the production process (measuring, drafting, copying). The traces discovered on the map of Cadzand (compass circles, perforations and incised meridians and bearings) are unprecedented in the history of cartography, can be directly related to the step-by-step method and drawings of Frisius, and provide evidence of the use of triangulation in the sixteenth century. Other traces, such as squaring grids, pounced dots or map inserts, further clarify the methods Pourbus used for compiling and copying maps.

Abstract
Towards the end of World War II, from January 1945, British/Canadian forces supplemented the use of topographical maps by compiling innovative specialist thematic maps to assist troops advancing eastwards from Belgium and the Netherlands into north-west Germany. 'Soil' maps were used to predict (1) cross-country trafficability and (2) potential airfield construction sites. Intensive bombardment of ground that was unexpectedly waterlogged as well as deliberate flooding by the enemy created mud which initially slowed cross-country movement. So too did determined German military resistance: the campaign included arguably the bloodiest operation for troops of 21st Army Group since their battle for Normandy in the summer of 1944. Significant maps, mostly at 1:100,000, are now preserved in England within the Shotton Archive at the Lapworth Museum of Geology, Birmingham. Together with expertise from geologists Major F.W. Shotton and Squadron Leader J.F. Kirkaldy, the maps contributed to terrain analysis that assisted with operational planning.

Abstract
Understanding the mobility of entities in geospatial data is important to many fields, ranging from the social sciences to epidemiology, economics or air traffic control. Visualizing such entities can be challenging as it requires preserving both their explicit properties (spatial trajectories) and their implicit properties (abstract attributes of those trajectories). An

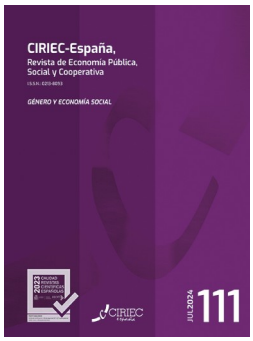
existing technique called origin–destination maps preserves both explicit and implicit properties of datasets, using the spatial nesting technique. In this paper, we aim at generalizing this technique beyond an origins-and-destinations dataset (2-attribute datasets), to explore multi-dimensional datasets (N-attribute datasets) with the nesting approach. We present an abstraction framework – we call Gridify – and an interactive open-source tool implementing this framework using several levels of nested maps. We report on several case studies representative of the types of dimensions found in geospatial datasets (quantitative, temporal, discrete, boolean), showing the applicability of this approach to achieve visual exploratory analysis tasks in various application domains.

Evaluating the Geometry of Objects in Cartographic Generalization with Hu's Invariants P. 69-88

Joanna Bac-Bronowicz, Krystian Kozioł, Andrzej Kwinta, Celso Augusto Guimarães Santos & Kamil Maciuk

Abstract

This paper presents a novel approach to assessing the geometry of objects using Hu's invariants in the context of cartographic generalization. The primary focus is to improve the generalization process and produce more readable and informative maps. The study demonstrates the applicability and effectiveness of the modified invariant moment M1* in evaluating regular shape similarity. Experiments, based on 24 shapes, exhibit greater stability in the results and reveal the high suitability of this moment in the investigation and classification of buildings, among other generalization processes. The efficiency of the proposed method is compared to previous generalization techniques, showing a significant improvement in the generalization process. In conclusion, this research contributes to the development of cartographic generalization methods by introducing the use of Hu's invariants for evaluating object geometry. This approach can improve the automation of map generalization processes and more effective communication of geographic information.



CIRIEC-ESPAÑA, revista de economía pública, social y cooperativa, ISSN 0213-8093
Número 111 (julio 2024)

Resumen
Las cooperativas -como entidades de economía social- priorizan los intereses de las personas socias y/o del fin social sobre el capital siguiendo el contenido de los Principios formulados por la Alianza Cooperativa Internacional (ACI). De entre estos últimos, y al objeto de este análisis, interesa prestar especial atención tanto al principio de gestión democrática y participativa en la entidad cooperativa, como al principio de igualdad, lo que en líneas generales implica la ausencia de desigualdad o discriminación. De este modo, con independencia del género, la persona que ostenta la condición de socia está facultada para participar en la actividad cooperativizada y en la estructura orgánica formando parte de los órganos sociales que la gestionan. La actual propuesta de reforma normativa -que pretende ser un impulso de la economía social (Anteproyecto de Ley Integral)- tiene como una de sus líneas prioritarias la adecuación del texto nacional a las vigentes exigencias y condicionantes de carácter económico y social. En consecuencia, y atendiendo al principio cooperativo de igualdad, se propugna la efectividad del mismo entre hombres y mujeres y, en particular, en cuanto a su significación en la gestión de la entidad cooperativa.

Resumen
El artículo propone una revisión de las leyes de cooperativas españolas con el objeto de identificar aquellas disposiciones que contienen medidas para favorecer la igualdad de género en la composición de los órganos de administración y dirección de estas sociedades. Teniendo en cuenta que la igualdad es uno de los valores esenciales de la identidad cooperativa, sería esperable que la concreción de este principio en el contexto del gobierno corporativo estuviera, cuando menos, al nivel del previsto para las sociedades capitalistas, especialmente las sociedades cotizadas con quienes las sociedades cooperativas comparten similar estructura corporativa. Sin embargo, el análisis evidencia que, si bien algunas leyes autonómicas destacan por haber incluido medidas significativas en esta materia de carácter genérico, societario, económico y de fomento, la mayoría sigue sin explicitar ningún compromiso con la igualdad de género. Por ello, se considera conveniente formular una serie de principios y recomendaciones sobre esta materia en la línea de los previstos en el Código de Buen Gobierno de Sociedades Cotizadas (CBGSC) a las que todo el movimiento debería adherirse sometiéndose al principio *comply or explain*. En este sentido, los Principios del Derecho Cooperativo Europeo (PECOL) podrían servir como inspiración.

Resumen
Este estudio examina la intención emprendedora de las mujeres en la Economía Social (ES) en Galicia, enfocándose en

los factores clave para cerrar la brecha de género en el emprendimiento y desarrollar modelos de negocio que respondan a las necesidades y preferencias femeninas. Utilizando datos de la encuesta GEM-Galicia, se analizan las motivaciones, barreras y dinámicas específicas del emprendimiento femenino en el ámbito de la ES. El estudio revela un desconocimiento significativo sobre estos modelos organizativos, junto con dificultades en financiación y burocracia. Los resultados destacan la necesidad urgente de mejorar la educación y el conocimiento sobre la ES entre las mujeres, y de implantar políticas públicas que proporcionen apoyo adaptado a sus necesidades. Además, se sugiere la promoción de modelos organizativos colaborativos y orientados al impacto social. Finalmente, se concluye que políticas específicas, incluyendo asesoramiento y formación adecuada, pueden fomentar un entorno más equitativo y propiciar un desarrollo empresarial centrado en las personas.

Empresas sociales e inclusión laboral de las mujeres con discapacidad. Las mujeres con discapacidad en los Centros Especiales de Empleo de la Comunidad Valenciana P. 97-130

Manuel Francisco Salinas Tomás, Elena Mut Montalva

Resumen

Las mujeres con discapacidad aúnan diversos factores de exclusión que les conduce a enfrentar situaciones específicas de desigualdad que, consecuentemente, les lleva a un déficit en su inclusión social y laboral.

Las empresas sociales, siempre a la vanguardia de favorecer el acceso al empleo de aquellos colectivos más desfavorecidos por la sociedad suponen una respuesta eficaz para su consecución. Los Centros Especiales de Empleo, en adelante CEE, como empresas sociales, suponen la vía más importante de acceso al trabajo para las personas con discapacidad en general, y para las mujeres con discapacidad en particular (Ortega, 2023; Salinas & Marhuenda, 2020).

El presente artículo justifica la necesidad de adoptar un análisis feminista para: superar la falta de visibilidad del colectivo, su percepción estereotipada y evidenciar las barreras que enfrentan para superar las exclusiones estructurales. Esta investigación presenta un panorama de la situación de las mujeres con discapacidad en los CEE de iniciativa social y, además, muestra datos comparativos entre la Comunitat Valenciana y el resto del estado español. Partimos de una metodología cuantitativa basada en el análisis de las bases de datos del INE y ODISMET.

Las conclusiones también muestran la necesidad de re-elaborar los indicadores de recogida de información que, en muchos casos, contribuyen a invisibilizar de una manera velada las necesidades reales de este colectivo.

Global care chains and empowerment through the social economy: women's participation in care cooperatives P. 131-160

Amaia Garcia-Azpuru, Julen Izagirre-Olaizola, Aitziber Etxezarreta-Etxarri, Liseth Díaz Molina

Resumen

El objetivo principal de este trabajo es analizar la participación de las mujeres en las cooperativas de cuidados como herramienta fundamental para su empoderamiento. Para ello, se discuten conceptos relacionados con la migración, la participación en la economía social y el empoderamiento de las mujeres. En una situación de crisis del sistema tradicional de cuidados en el Norte Global, la migración femenina extranjera viene a resolver parcialmente el problema. Sin embargo, se enfrentan a condiciones de vulnerabilidad y precariedad. Surgen cadenas globales de cuidados que conllevan elevados costes familiares y personales para el eslabón más débil.

En este contexto, este trabajo desarrolla un análisis cualitativo basado en entrevistas en profundidad a trabajadoras de tres cooperativas de cuidados del País Vasco. El objetivo es, a través de un análisis temático exploratorio, poner de manifiesto la forma en que la economía social, a través de las cooperativas de cuidados, puede ser una herramienta fundamental para el empoderamiento de las mujeres (migrantes o no) que se dedican al sector de los cuidados.

El análisis de la información revela que la cooperativa es una forma adecuada de fomentar la participación de las mujeres y esto, a su vez, conduce a un aumento de la confianza, la autoestima y, en general, el empoderamiento de

las mujeres. Sin embargo, también se identifican dificultades y barreras adicionales para las mujeres extranjeras en situación irregular.

Escuchando las mujeres: cooperativas y asociaciones sociales y solidarias como red de apoyo, un ejemplo brasileño P. 161-190

Julia da Silva Gutierrez Ruiz, Leandro Pereira Morais

Resumen

En Brasil, muchas mujeres son víctimas de distintas formas de violencia doméstica, que puede ser física, psicológica, moral, sexual y patrimonial. Las mujeres víctimas de violencia doméstica tienden a buscar personas cercanas para pedir ayuda para salir de la situación, por lo tanto, una red de apoyo informal puede ser una herramienta importante para que las mujeres se sientan protegidas y pidan ayuda a las organizaciones públicas. Dado que las empresas económicas solidarias (EES) formadas solamente por mujeres son un espacio para que las mujeres se fortalezcan mutuamente, donde socializan y pueden crear vínculos, también pueden servir como red de apoyo informal para las mujeres que las forman, además de fomentar su autonomía financiera. El objetivo de este artículo es identificar si las EES pueden formar parte de una red de apoyo a las mujeres, desde la perspectiva de las mujeres que las forman. Para ello, se investigó tres EES de mujeres en el municipio brasileño de Araraquara, que tiene una importante historia de movimientos sociales y políticas públicas dirigidas a la economía social y solidaria. A partir de las entrevistas semiestructuradas realizadas a 11 mujeres de tres de estos EES, fue posible obtener resultados que explican lo que ocurre en el campo empírico en relación con la teoría, como que las mujeres se sienten cómodas y confían unas en otras, pero al mismo tiempo no identifican la violencia patrimonial como violencia.

¿Es más frágil el techo de cristal en la Economía Social? Un análisis en cooperativas y sociedades laborales españolas P. 191-225

Belén Castro Núñez, Lidia de Castro Romero, Víctor Martín Barroso, Rosa Santero-Sánchez

Resumen

Las cooperativas y sociedades laborales, como entidades representativas de la Economía Social, son empresas que priorizan a las personas por encima del capital y, estos valores se reflejan en plantillas más igualitarias e inclusivas. Los análisis comparados de este tipo de empresas frente a empresas mercantiles muestran menores brechas de género, tanto en la participación en el empleo como en las condiciones laborales. Este trabajo tiene como objetivo calcular la probabilidad de acceder a los puestos de mayor responsabilidad en las empresas y ver si en estas entidades de economía social, las mujeres tienen mayores probabilidades de acceso que en las empresas mercantiles, concluyendo así que experimentan menores obstáculos y considerando que el techo de cristal es más frágil. A partir de datos registrales de seguridad social y del análisis de trayectorias laborales en una década, las distintas estimaciones realizadas ponen de manifiesto que las trabajadoras de cooperativas y sociedades laborales presentan una mayor probabilidad de alcanzar el grupo de mayor cotización que los hombres. En el periodo analizado, estas empresas han favorecido la incorporación femenina a los puestos de decisión y no han encontrado techo de cristal. Esta situación se produce principalmente para el grupo de personas con edades por encima de los 30 años.

Liderazgo femenino en los consejos y respuesta corporativa a las iniciativas de sostenibilidad: Un caso exploratorio en cooperativas agroalimentarias P. 227-259

Elia García-Martí, Mª Jesús Hernández-Ortiz, Mª del Carmen Ruiz-Jiménez, Cristina Pedrosa-Ortega, Rocío Martínez-Jiménez

Resumen

Actualmente existe un gran interés por la participación de la mujer en ámbitos de toma de decisiones, debido a que diferentes estudios demuestran que la presencia femenina en la dirección influye en la mejora de la sostenibilidad de la empresa.

Este trabajo contribuye a mejorar la comprensión de las características del liderazgo femenino en cooperativas agroalimentarias, a través de un estudio de casos de referencia en España. El objetivo ha sido analizar el proceso de la mujer en el acceso al consejo rector, las características de su liderazgo y la influencia que su liderazgo tiene en la sostenibilidad empresarial, medida a través de sus tres componentes ESG (Environmetal, Social and Governance).

Los resultados indican que la mujer accede al consejo rector tras demostrar sus capacidades profesionales ante los socios. Su liderazgo se puede definir como transformacional y se apoya en la cultura cooperativa. Además, se pone de relieve que existe una relación positiva entre la mujer en puestos de liderazgo y la mejora de la sostenibilidad en cooperativas agroalimentarias.

Análisis de la transformación digital en las cooperativas agroalimentarias desde la perspectiva de género P. 261-303

Carmen Guzmán, Francisco J. Santos, Pedro Ahumada

Resumen

La transformación digital (TD) implica la aparición de nuevos modelos de negocio basados en el uso generalizado de las tecnologías digitales. La TD es necesaria para mejorar la productividad y el acceso a los mercados; sin embargo, desde la perspectiva de género, existe una brecha digital. En esta investigación, el objetivo es estudiar la TD desde la perspectiva de género en un sector específico, el agroalimentario, y en un tipo de entidad específico de la economía social que es fundamental en las zonas rurales, las cooperativas agroalimentarias. En estas empresas sigue habiendo una brecha de género, especialmente debido a la menor proporción de mujeres en los equipos directivos y, además, presentan un atraso en su TD respecto a otras empresas. En concreto, se analiza la TD en estas cooperativas tanto globalmente como mediante el análisis de cada una de sus dimensiones: infraestructuras, productos, organización, procesos y clientes. Para ello, se utilizan los datos de una encuesta a una muestra de cooperativas agroalimentarias extremeñas, región con una fuerte especialización en el sector agroalimentario y donde las cooperativas desempeñan una importante función económica, social y medioambiental. Los resultados muestran que la presencia de mujeres en la presidencia de estas cooperativas influye positivamente en la presencia de mujeres en sus equipos directivos. Asimismo, no existen diferencias significativas en la TD global de estas cooperativas respecto al factor género, aunque sí en las dimensiones específicas de “clientes” y “procesos”.

Work-family conflict in social economy organisations. Individual differences in the employees’ demographic profile P. 305-327

Esther Villajos, Amalia Pérez-Nebra, Maite Legarra, Eunate Elio

Resumen

Los cambios y la diversidad en el perfil demográfico de la población activa española han despertado un gran interés, especialmente en la gestión de recursos humanos. Examinar el intrincado equilibrio que los individuos y las organizaciones establecen entre las responsabilidades familiares, de género y laborales resulta especialmente crucial, sobre todo en el contexto de las familias con hijos. El conflicto entre trabajo y familia puede originarse en cualquiera de los dos ámbitos. Así pues, la dirección del conflicto adquiere relevancia tanto a nivel personal como organizativo, para entender si el trabajo interfiere con la familia (WFC), o la familia interfiere con el trabajo (FWC). Sin embargo, en las organizaciones de la economía social estos conflictos pueden manifestarse de forma diferente, dado que este tipo de organizaciones tienden a ser más horizontales, sociales, femeninas y participativas. Esto significa que el conflicto entre familia y trabajo podría ser menos perjudicial. Teniendo esto en cuenta, nuestro objetivo era analizar si los hijos, en función de su edad, influyen en el conflicto trabajo-familia, y evaluar las posibles diferencias de género. Los resultados revelaron algunas diferencias en cuanto a la edad de los hijos en el conflicto trabajo-familia y familia-trabajo entre mujeres y hombres. A lo largo del artículo se discuten las implicaciones prácticas y teóricas.

The right to work-life balance in Portuguese cooperatives. A legal and empirical analysis P. 329-360

Deolinda Meira, Conceição Castro, Sofia Antunes

Resumen

Este artículo tiene como objetivo evaluar si las cooperativas son el entorno ideal para fomentar el equilibrio entre el trabajo y la vida personal, un derecho cuya protección se ha desarrollado en estrecha conexión con la promoción de la igualdad de género. Con este fin, pretendemos responder a las siguientes preguntas: (i) ¿Cuáles son las dimensiones principales del derecho al equilibrio entre el trabajo y la vida personal, y cuál es la relevancia de proteger este derecho para promover la igualdad de género? (ii) ¿En qué medida puede el régimen legal de las cooperativas promover el derecho al equilibrio entre el trabajo y la vida personal y, en consecuencia, la igualdad de género? (iii) ¿Cuáles son los principales facilitadores y obstáculos para el equilibrio entre el trabajo y la vida personal en las cooperativas desde la perspectiva de mujeres y hombres?

Desde un punto de vista metodológico, nuestro estudio se basa en una revisión de la literatura, la legislación y un estudio empírico. El análisis empírico se basó en los resultados de una encuesta de cuestionario, que arrojó 414 respuestas válidas de empleados de SEEs, incluyendo 62 de cooperativas. Los métodos estadísticos seguidos fueron estadísticas descriptivas, pruebas no paramétricas y pruebas post hoc. Los hallazgos arrojan luz sobre los desafíos específicos enfrentados por los empleados en las cooperativas con respecto al equilibrio entre el trabajo y la vida personal, y resaltan la importancia de canales de comunicación abiertos y de apoyo en el lugar de trabajo y la flexibilidad en la gestión del tiempo. Los resultados también sugieren que, según la percepción de los empleados, las cooperativas son las que promueven más facilitadores y que su cultura y clima organizacional son propicios para fomentar un equilibrio saludable entre el trabajo y la vida personal. Los resultados observados pueden atribuirse principalmente a la gestión democrática y participativa inherente a las estructuras cooperativas.

TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 32, number 3 (september 2023)

Abstract
Matrix-valued data arise in many applications. In this paper, we consider the setting where one collects both a matrix-valued data $\mathbf{Y} \in R^{pxq}$ and a generic scalar X that can be continuous, discrete or categorical. Since the rows and columns of **Y** often have specific meanings in practice, it is interesting to make statistical inferences on the significance of rows and columns of **Y**. In this paper, by taking into account the background effect, we propose a new measure on significance of rows and columns based on an additive model. The point estimates, hypothesis testings and confidence intervals of the significance of a given row or column of **Y** are considered. Moreover, a procedure is proposed to select significant rows and columns. Our method is applicable to both p and q being much larger than sample size n. Simulation results and real data analysis demonstrate the effectiveness of the proposed method.

Abstract
This paper proposes a new test for the comparison of conditional quantile curves when the outcome of interest, typically a duration, is subject to right censoring. The test can be applied both in the case of two independent samples and for paired data, and can be used for the comparison of quantiles at a fixed quantile level, a finite set of levels or a range of quantile levels. The asymptotic distribution of the proposed test statistics is obtained both under the null hypothesis and under local alternatives. We describe a bootstrap procedure in order to approximate the critical values and present the results of a simulation study, in which the performance of the tests for small and moderate sample sizes is studied and compared with the behavior of alternative tests. Finally, we apply the proposed tests on a data set concerning diabetic retinopathy.

Abstract
In reliability engineering, the relevation model can be adopted to characterize the performance of redundancy allocation for coherent systems. In this paper, we investigate the allocation problems of relevations for two nodes in a coherent system with independent components for enhancing system reliability. We first investigate the optimal allocation policy of two relevations for two nodes of the system under certain conditions. As a special setting of the relevation, we further discuss optimal allocation strategies for a batch of minimal repairs allocated to two components of the coherent system by applying the useful tool of majorization order. Sufficient conditions are established in terms of structural relationships between the components induced by minimal cut or path sets and the reliabilities of components and relevations. Some numerical examples are provided as illustrations. A real application in aircraft

indicator lights systems is also presented to show the availability of our results.

Novel specification tests for synchronous additive concurrent model formulation based on martingale difference divergence P. 908-941

Laura Freijeiro-González ; Manuel Febrero-Bande ; Wenceslao González-Manteiga

Abstract

This paper presents new specification tests for a general synchronous additive concurrent model formulation. As a novelty, our proposal does not require a preliminary model or error structure estimation. No tuning parameters are involved either. We develop a suitable test statistic using the martingale difference divergence coefficient. As a result, this statistic measures the departure from the conditional mean independence in the concurrent model framework, considering the information of all observed time instants. In particular, global as well as partial dependence tests are introduced. Then, we allow one to quantify the effect of a group of covariates or to apply covariates selection one by one. We obtain its asymptotic distribution under the null and propose a bootstrap algorithm to compute the p-values in practice. Through simulations, we illustrate our method, and its performance is compared to existing competitors. In addition, we use this in the analysis of three real datasets related to gait data, flu activity, and casual bike rentals.

Level sets of depth measures in abstract spaces P. 942-957

A. Cholaquidis ; R. Fraiman ; L. Moreno

Abstract

The lens depth of a point has been recently extended to general metric spaces, which is not the case for most depths. It is defined as the probability of being included in the intersection of two random balls centred at two random points X and Y, with the same radius $d(X, Y)$. We prove that, on a separable and complete metric space, the level sets of the empirical lens depth based on an iid sample, converge in the Painlevé–Kuratowski sense, to its population counterpart. We also prove that, restricted to compact sets, the empirical level sets and their boundaries are consistent estimators, in Hausdorff distance, of their population counterparts, and analyse two real-life examples.

Estimating weak periodic vector autoregressive time series P. 958-997

Yacouba Boubacar Maïnassara ; Eugen Ursu

Abstract

This article develops the asymptotic distribution of the least squares estimator of the model parameters in periodic vector autoregressive time series models (hereafter PVAR) with uncorrelated but dependent innovations. When the innovations are dependent, this asymptotic distributions can be quite different from that of PVAR models with independent and identically distributed (iid for short) innovations developed (Ursu and Duchesne in J Time Ser Anal 30:70–96, 2009). Modified versions of the Wald tests are proposed for testing linear restrictions on the parameters. These asymptotic results are illustrated by Monte Carlo experiments. An application to a bivariate real financial data is also proposed.

Hypothesis testing in adaptively sampled data: ART to maximize power beyond *iid* sampling P. 998-1037

Dae Woong Ham ; Jiaze Qiu

Abstract

Testing whether a variable of interest affects the outcome is one of the most fundamental problems in statistics and is often the main scientific question of interest. To tackle this problem, the conditional randomization test (CRT) is widely used to test the independence of variable(s) of interest (X) with an outcome (Y) holding other variable(s) (Z) fixed. The CRT uses "Model-X" inference framework that relies solely on the *iid* sampling of (X, Z) to produce exact finite-sample p values that are constructed using any test statistic. We propose a new method, the *adaptive*

randomization test (ART), that tackles the same independence problem while allowing the data to be adaptively sampled. Like the CRT, the ART relies solely on knowing the (adaptive) sampling distribution of (X, Z). Although the ART allows practitioners to flexibly design and analyze adaptive experiments, the method itself does not guarantee a powerful adaptive sampling procedure. For this reason, we show substantial power gains obtained from adaptively sampling compared to the typical *iid* sampling procedure in a multi-arm bandit setting and an application in conjoint analysis. We believe that the proposed adaptive procedure is successful because it takes arms that may initially look like "fake" signals due to random chance and stabilizes them closer to "null" signals and samples more/less from signal/null arms.

Reliability and optimal replacement policy for a generalized mixed shock model P. 1038-1054
Murat Ozkut

Abstract
A generalized mixed shock model, which mixes two run shock models, is developed and analyzed. According to the model, the system subject to both internal degradation and external shocks fails upon the occurrence of k_1 consecutive shocks whose magnitude is between predefined critical values of d_1 and d_2 such that $d_1 < d_2$, or k_2 consecutive shocks whose magnitude is above d_2. The system's reliability, mean time to failure, and mean residual lifetime are all calculated under the assumption that the lifetime of the system due to internal wear and external shock arrival times follows a phase-type distribution. The best policy for replacement is also discussed. There are also graphical representations and numerical examples for the proposed model, in which both lifetime distribution of internal degradation and the interarrival periods between external shocks follow the Erlang distribution.

Robust and efficient estimation of nonparametric generalized linear models P. 1055-1078
Ioannis Kalogridis ; Gerda Claeskens ; Stefan Van Aelst

Abstract
Generalized linear models are flexible tools for the analysis of diverse datasets, but the classical formulation requires that the parametric component is correctly specified and the data contain no atypical observations. To address these shortcomings, we introduce and study a family of nonparametric full-rank and lower-rank spline estimators that result from the minimization of a penalized density power divergence. The proposed class of estimators is easily implementable, offers high protection against outlying observations and can be tuned for arbitrarily high efficiency in the case of clean data. We show that under weak assumptions, these estimators converge at a fast rate and illustrate their highly competitive performance on a simulation study and two real-data examples.

A general class of shock models with dependent inter-arrival times P. 1079-1105
Dheeraj Goyal ; Nil Kamal Hazra ; Maxim Finkelstein

Abstract
We introduce and study a general class of shock models with dependent inter-arrival times of shocks that occur according to the homogeneous Poisson generalized gamma process. A lifetime of a system affected by a shock process from this class is represented by the convolution of inter-arrival times of shocks. This class contains many popular shock models, namely the extreme shock model, the generalized extreme shock model, the run shock model, the generalized run shock model, specific mixed shock models, etc. For systems operating under shocks, we derive and discuss the main reliability characteristics (namely the survival function, the failure rate function, the mean residual lifetime function and the mean lifetime) and study relevant stochastic comparisons. Finally, we provide some numerical examples and illustrate our findings by the application that considers an optimal mission duration policy.

Nonparametric tests for semiparametric regression models P. 1106-1130
Federico Ferraccioli ; Laura M. Sangalli ; Livio Finos

Abstract
Semiparametric regression models have received considerable attention over the last decades, because of their flexibility and their good finite sample performances. Here we propose an innovative nonparametric test for the linear part of the models, based on random sign-flipping of an appropriate transformation of the residuals, that exploits a spectral decomposition of the residualizing matrix associated with the nonparametric part of the model. The test can be applied to a vast class of extensively used semiparametric regression models with roughness penalties, with nonparametric components defined over one-dimensional, as well as over multi-dimensional domains, including, for instance, models based on univariate or multivariate splines. We prove the good asymptotic properties of the proposed test. Moreover, by means of extensive simulation studies, we show the superiority of the proposed test with respect to current parametric alternatives, demonstrating its excellent control of the Type I error, accompanied by a good power, even in challenging data scenarios, where instead current parametric alternatives fail.

TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 33, number 3 (september 2024)

Abstract
Modern data collection in many data paradigms, including bioinformatics, often incorporates multiple traits derived from different data types (i.e., platforms). We call this data multi-block, multi-view, or multi-omics data. The emergent field of data integration develops and applies new methods for studying multi-block data and identifying how different data types relate and differ. One major frontier in contemporary data integration research is methodology that can identify partially shared structure between sub-collections of data types. This work presents a new approach: Data Integration Via Analysis of Subspaces (DIVAS). DIVAS combines new insights in angular subspace perturbation theory with recent developments in matrix signal processing and convex–concave optimization into one algorithm for exploring partially shared structure. Based on principal angles between subspaces, DIVAS provides built-in inference on the results of the analysis, and is effective even in high-dimension-low-sample-size (HDLSS) situations.

Abstract
Modern data collection in many data paradigms, including bioinformatics, often incorporates multiple traits derived from different data types (i.e., platforms). We call this data multi-block, multi-view, or multi-omics data. The emergent field of data integration develops and applies new methods for studying multi-block data and identifying how different data types relate and differ. One major frontier in contemporary data integration research is methodology that can identify partially shared structure between sub-collections of data types. This work presents a new approach: Data Integration Via Analysis of Subspaces (DIVAS). DIVAS combines new insights in angular subspace perturbation theory with recent developments in matrix signal processing and convex–concave optimization into one algorithm for exploring partially shared structure. Based on principal angles between subspaces, DIVAS provides built-in inference on the results of the analysis and is effective even in high-dimension-low-sample-size (HDLSS) situations.

Abstract
An ongoing “replication crisis” calls into question scientific discoveries across a variety of disciplines ranging from life to social sciences. Replication studies aim to investigate the validity of findings in published research, and try to assess whether the latter are statistically consistent with those in the replications. While the majority of replication projects are based on a single experiment, multiple independent replications of the same experiment conducted simultaneously at different sites are becoming more frequent. In connection with these types of projects, we deal with testing heterogeneity among sites; specifically, we focus on sample size determination suitable to deliver compelling evidence once the experimental data are gathered.

On variability of the mean remaining lifetime at random age P. 717-730

Majid Asadi ; Maxim Finkelstein

Abstract

In this short communication, we discuss the remaining lifetime and the mean remaining lifetime (MRL) of an item with a random age. We show that the MRL at random age is closely related to some well-known variability measures. First, we provide a decomposition result showing that the MRL at random age, similar to other variability measures, has a covariance representation. Under the proportional hazards (PH) model, we show that the MRL, depending on the parameter of proportionality, subsumes the Gini's mean difference and the cumulative residual entropy as special cases. It is also shown that, under the PH model, the MRL can be expressed via the equilibrium distribution and the mean number of events in the generalized Pólya process.

A copula formulation for multivariate latent Markov models P. 731-751

Alfonso Russo ; Alessio Farcomeni

Abstract

We specify a general formulation for multivariate latent Markov models for panel data, where outcomes are possibly of mixed-type (categorical, discrete, continuous). Conditionally on a time-varying discrete latent variable and covariates, the joint distribution of outcomes simultaneously observed is expressed through a parametric copula. We therefore do not make any conditional independence assumption. The observed likelihood is maximized by means of an expectation–maximization algorithm. In a simulation study, we argue how modeling the residual contemporary dependence might be crucial in order to avoid bias in the parameter estimates. We illustrate through an original application to assessment of poverty through direct and indirect indicators in a cohort of Italian households.

The orthogonal skew model: computationally efficient multivariate skew-normal and skew-t distributions with applications to model-based clustering P. 752-785

Ryan P. Browne ; Jeffrey L. Andrews

Abstract

We introduce a parameterization for the multivariate skew normal and skew-t distributions, which enforces an orthogonal structure on the skewness parameter. This approach provides substantial benefits in computational efficiency during parameter estimation, resulting in a model which strikes an excellent balance between flexibility and model-fitting feasibility. We illustrate this primarily through implementing the proposed distributions in a mixture model-based clustering framework. We compare to competing skew distributions via both simulated and real data analyses, reporting both computation time and model-fit metrics.

Two-step semiparametric empirical likelihood inference from capture–recapture data with missing covariates P. 786-808

Yang Liu ; Yukun Liu ; Riquan Zhang

Abstract

Missing covariates are not uncommon in capture–recapture studies. When covariate information is missing at random in capture–recapture data, an empirical full likelihood method has been demonstrated to outperform conditional-likelihood-based methods in abundance estimation. However, the fully observed covariates must be discrete, and the method is not directly applicable to continuous-time capture–recapture data. Based on the Binomial and Poisson regression models, we propose a two-step semiparametric empirical likelihood approach for abundance estimation in the presence of missing covariates, regardless of whether the fully observed covariates are discrete or continuous. We show that the maximum semiparametric empirical likelihood estimators for the underlying parameters and the abundance are asymptotically normal, and more efficient than the counterpart for a completely known non-missingness probability. After scaling, the empirical likelihood ratio test statistic for abundance follows a limiting chi-square distribution with one degree of freedom. The proposed approach is further extended to one-inflated count

regression models, and a score-like test is constructed to assess whether one-inflation exists among the number of captures. Our simulation shows that, compared with the previous method, the proposed method not only performs better in correcting bias, but also has a more accurate coverage in the presence of fully observed continuous covariates, although there may be a slight efficiency loss when the fully observed covariates are only discrete. The performance of the new method is illustrated by analyses of the yellow-bellied prinia data and the rana pretiosa data.

Multiple change point detection for high-dimensional data P. 809-846

Wenbiao Zhao ; Lixing Zhu ; Falong Tan

Abstract

This research investigates the detection of multiple change points in high-dimensional data without particular sparse or dense structure, where the dimension can be of exponential order in relation to the sample size. The estimation approach proposed employs a signal statistic based on a sequence of signal screening-based local U-statistics. This technique avoids costly computations that exhaustive search algorithms require and mitigates false positives, which hypothesis testing-based methods need to control. Consistency of estimation can be achieved for both the locations and number of change points, even when the number of change points diverges at a certain rate as the sample size increases. Additionally, the visualization nature of the proposed approach makes plotting the signal statistic a useful tool to identify locations of change points, which distinguishes it from existing methods in the literature. Numerical studies are performed to evaluate the effectiveness of the proposed technique in finite sample scenarios, and a real data analysis is presented to illustrate its application.

Testing covariance structures belonging to a quadratic subspace under a doubly multivariate model P. 847-876

Katarzyna Filipiak ; Mateusz John ; Yuli Liang

Abstract

A hypothesis related to the block structure of a covariance matrix under the doubly multivariate normal model is studied. It is assumed that the block structure of the covariance matrix belongs to a quadratic subspace, and under the null hypothesis, each block of the covariance matrix also has a structure belonging to some quadratic subspace. The Rao score and the likelihood ratio test statistics are derived, and the exact distribution of the likelihood ratio test is determined. Simulation studies show the advantage of the Rao score test over the likelihood ratio test in terms of speed of convergence to the limiting chi-square distribution, while both proposed tests are competitive in terms of their power. The results are applied to both simulated and real-life example data.

Privacy-preserving parametric inference for spatial autoregressive model P. 847-896

Zhijian Wang ; Yunquan Song

Abstract

Spatial regression models are important tools in dealing with spatially dependent data and are widely used in many fields such as spatial econometric and regional science. When the spatial data contain sensitive information, the privacy of the data will be compromised along with the release of the analysis if appropriate privacy-preserving measures are not in place. In this paper, we study the privacy-preserving parametric inference for the spatial autoregressive model and propose corresponding differentially private algorithm. We construct a differentially private spatial autoregression framework that takes graph data into account. We improve the functional mechanism to be more accurate under the same degree of privacy protection. Theoretical analysis establishes both the privacy guarantees of the algorithm and the asymptotic normality of the estimation. Simulation and real data studies show improvements of our approach.

Abstract
We consider a semiparametric partly linear model identified by instrumental variables. We propose an estimation method that does not smooth on the instruments and we extend the Landweber–Fridman regularization scheme to the estimation of this semiparametric model. We then show the asymptotic normality of the parametric estimator and obtain the convergence rate for the nonparametric estimator. Our estimator that does not smooth on the instruments coincides with a typical estimator that does smooth on the instruments but keeps the respective bandwidth fixed as the sample size increases. We propose a data driven method for the selection of the regularization parameter, and in a simulation study we show the attractive performance of our estimators.

Abstract
Sufficient dimension reduction is commonly performed to achieve data reduction and help data visualization. Its main goal is to identify functions of the predictors that are smaller in number than the predictors and contain the same information as the predictors for the response. In this paper, we are concerned with the linear functions of the predictors, which determine a central subspace that preserves sufficient information about the conditional distribution of a response given covariates. Many methods have been developed in the literature for the estimation of the central subspace. However, most of the existing sufficient dimension reduction methods are sensitive to outliers and require some strict restrictions on both covariates and response. To address this, we propose a novel dependence measure, rank divergence, and develop a rank divergence-based sufficient dimension reduction approach. The new method only requires some mild conditions on the covariates and response and is robust to outliers or heavy-tailed distributions. Moreover, it applies to both discrete or categorical covariates and multivariate responses. The consistency of the resulting estimator of the central subspace is established, and numerical studies suggest that it works well in practical situations.

Abstract
In statistical diagnostics, detecting influential observations is pivotal for assessing model fitting. To address parameter restrictions while maintaining necessary properties, the softplus INGARCH model has emerged as an alternative to the INGARCH model and its variants. This paper delves into statistical diagnostics within the softplus INGARCH model using local influence analysis, establishing a framework encompassing first-order diagnostics, second-order diagnostics and stepwise diagnostics. Additionally, we focus on perturbation schemes, refining conventional approaches and offering modifications. To demonstrate the effectiveness and suitability of our proposed methodology, particularly with the inclusion of stepwise diagnostics, we analyze two simulated datasets and two real-world examples. Compared to traditional methods, our approach adeptly handles potential issues such as the “masking effect” and “smearing effect” without necessitating complex calculations.

TEST : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1133-0686
Volume 33, number 4 (december 2024)

Abstract
Prediction, in regression and classification, is one of the main aims in modern data science. When the number of predictors is large, a common first step is to reduce the dimension of the data. Sufficient dimension reduction (SDR) is a well-established paradigm of reduction that keeps all the relevant information in the covariates X that is necessary for the prediction of Y. In practice, SDR has been successfully used as an exploratory tool for modeling after estimation of the sufficient reduction. Nevertheless, even if the estimated reduction is a consistent estimator of the population, there is no theory supporting this step when nonparametric regression is used in the imputed estimator. In this paper, we show that the asymptotic distribution of the nonparametric regression estimator remains unchanged whether the true SDR or its estimator is used. This result allows making inferences, for example, computing confidence intervals for the regression function, thereby avoiding the curse of dimensionality.

Abstract
We propose a goodness-of-fit test for a class of count time series models with covariates which includes the Poisson autoregressive model with covariates (PARX) as a special case. The test criteria are derived from a specific characterization for the conditional probability generating function, and the test statistic is formulated as a L_2 weighting norm of the corresponding sample counterpart. The asymptotic properties of the proposed test statistic are provided under the null hypothesis as well as under specific alternatives. A bootstrap version of the test is explored in a Monte–Carlo study and illustrated on a real data set on road safety.

Abstract
Single-index models are important and popular semiparametric models, as they can handle the problem of the “curse of dimensionality” and enjoy the flexibility of nonparametric modeling and the interpretability of parametric modeling. Most existing methods for single-index models are sensitive to outliers or heavy-tailed distributions because they use the least squares criterion. An oracle-efficient M-estimator is proposed for single-index models, and a smooth simultaneous confidence band is constructed by treating the index coefficients as nuisance parameters. Under general assumptions it is shown that the M-estimator for the nonparametric link function, based on any $\sqrt{n}$-consistent coefficient index parameter estimators, is oracle-efficient. This means that it is uniformly as efficient as the infeasible one obtained by M-regression using the true single-index coefficient parameters. As a result, the asymptotic distribution of the maximal deviation between the M-type kernel estimator and the true link function is

derived, and an asymptotically accurate simultaneous confidence band is established as a global inference tool for the link function. The proposed method generalizes the desirable uniform convergence property of ordinary least squares to the M-estimation. Meanwhile, it is a general approach that allows any $\sqrt{n}$-consistent coefficient parameter estimators to be applied in the procedure to make global inferences for the link function. Simulation studies with commonly encountered sample sizes are reported to support the theoretical findings. These numerical results show certain desirable robustness properties against heavy-tailed errors and outliers. As an illustration, the proposed method is applied to the analysis of a car purchasing dataset.

Conformal link prediction for false discovery rate control P. 1062-1083

Ariane Marandon

Abstract

Most link prediction methods return estimates of the connection probability of missing edges in a graph. Such output can be used to rank the missing edges from most to least likely to be a true edge, but does not directly provide a classification into true and nonexistent. In this work, we consider the problem of identifying a set of true edges with a control of the false discovery rate (FDR). We propose a novel method based on high-level ideas from the literature on conformal inference. The graph structure induces intricate dependence in the data, which we carefully take into account, as this makes the setup different from the usual setup in conformal inference, where data exchangeability is assumed. The FDR control is empirically demonstrated for both simulated and real data.

Optimal subsampling for L*p*-quantile regression via decorrelated score P. 1084-1104

Xing Li ; Yujing Shao ; Lei Wang

Abstract

To balance robustness of quantile regression and effectiveness of expectile regression, we consider L*p*-quantile regression models with large-scale data and develop a unified optimal subsampling method to downsize the data volume and reduce computational burden. For low-dimensional L*p*-quantile regression models, two optimal subsampling probabilities based on the A- and L-optimality criteria are firstly proposed. For the preconceived low-dimensional parameter in high-dimensional L*p*-quantile regression models, a novel optimal subsampling decorrelated score function is proposed to mitigate the effect from nuisance parameter estimation and then two optimal decorrelated score subsampling probabilities are provided. The asymptotic properties of two optimal subsample estimators are established. The finite-sample performance of the proposed estimators is studied through simulations, and an application to Beijing Air Quality Dataset is also presented.

Marginal analysis of count time series in the presence of missing observations P. 1105-1128

Simon Nik

Abstract

Time series in real-world applications often have missing observations, making typical analytical methods unsuitable. One method for dealing with missing data is the concept of amplitude modulation. While this principle works with any data, here, missing data for unbounded and bounded count time series are investigated, where tailor-made dispersion and skewness statistics are used for model diagnostics. General closed-form asymptotic formulas are derived for such statistics with only weak assumptions on the underlying process. Moreover, closed-form formulas are derived for the popular special cases of Poisson and binomial autoregressive processes, always under the assumption that missingness occurs. The finite-sample performances of the considered asymptotic approximations are analyzed with simulations. The practical application of the corresponding dispersion and skewness tests under missing data is demonstrated with three real data examples.

Jackknife empirical likelihood for the correlation coefficient with additive distortion measurement errors P. 1129-1159

Da Chen ; Linlin Dai ; Yichuan Zhao

Abstract

The correlation coefficient is fundamental in advanced statistical analysis. However, traditional methods of calculating correlation coefficients can be biased due to the existence of confounding variables. Such confounding variables could act in an additive or multiplicative fashion. To study the additive model, previous research has shown residual-based estimation of correlation coefficients. The powerful tool of empirical likelihood (EL) has been used to construct the confidence interval for the correlation coefficient. However, the methods so far only perform well when sample sizes are large. With small sample size situations, the coverage probability of EL, for instance, can be below 90% at confidence level 95%. On the basis of previous research, we propose new methods of interval estimation for the correlation coefficient using jackknife empirical likelihood, mean jackknife empirical likelihood and adjusted jackknife empirical likelihood. For better performance with small sample sizes, we also propose mean adjusted empirical likelihood. The simulation results show the best performance with mean adjusted jackknife empirical likelihood when the sample sizes are as small as 25. Real data analyses are used to illustrate the proposed approach.

Extended Hotelling T^2 test in distributed frameworks P. 1160-1179

Bin Du ; Xiumin Liu ; Junlong Zhao

Abstract

Hypothesis test for a mean vector is a classical problem in data analysis but has been highly underinvestigated in distributed frameworks where samples of size n are located on k local sites. This paper focuses on the one-sample mean test, proposing synthesized test statistics with a much lower communication cost than the centralized Hotelling T^2 test. For the homogeneous case, where data on different local sites are independent and identically distributed, the efficiency of our proposed test is comparable to that of the centralized one, and much better than the test constructed from the divide and conquer method. Besides, three heterogeneous cases are considered, where the distributions of the data on local sites can be different. Heterogeneous cases are much more challenging because the local sample means and covariance matrices may be inconsistent estimators. We construct communication-efficient testing procedures for heterogeneous cases, and the power of the proposed test statistics is comparable to that of the centralized one under some conditions. Simulation results verify the effectiveness of the proposed testing procedures.

Modeling paired binary data by a new bivariate Bernoulli model with flexible beta kernel correlation P. 1180-1224

Xun-Jian Li ; Shuang Li ; Jianhua Shi

Abstract

Paired binary data often appear in studies of subjects with two sites such as eyes, ears, lungs, kidneys, feet and so on. Three popular models [i.e., (Rosner in Biometrics 38:105-114, 1982) R model, (Dallal in Biometrics 44:253-257, 1988) model and (Donner in Biometrics 45:605-661, 1989) model] were proposed to fit such twin data by considering the intra-person correlation. However, Rosner's R model can only fit the twin data with an increasing correlation coefficient, Dallal's model may incur the problem of over–fitting, while Donner's model can only fit the twin data with a constant correlation. This paper aims to propose a new *bivariate Bernoulli model with flexible beta kernel correlation* (denoted by Bernoulli_2^{bk}) for fitting the paired binary data with a wide range of group–specific disease probabilities. The correlation coefficient of the Bernoulli_2^{bk} model could be increasing, or decreasing, or unimodal, or convex with respect to the disease probability of one eye. To obtain the *maximum likelihood estimates* (MLEs) of parameters, we develop a series of *minorization–maximization* (MM) algorithms by constructing four surrogate functions with closed–form expressions at each iteration of the MM algorithms. Simulation studies are conducted, and two real datasets are analyzed to illustrate the proposed model and methods.

Nonparametric conditional survival function estimation and plug-in bandwidth selection with multiple covariates P. 1225-1257

Dimitrios Bagkavos ; Montserrat Guillen ; Jens P. Nielsen

Abstract

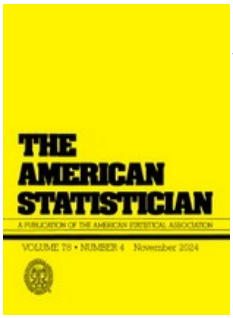
The present research provides two methodological advances, simulation evidence and a real data analysis, all contributing to the area of local linear survival function estimation and bandwidth selection. The first contribution is the development of a double smoothed local linear survival function estimator which admits an arbitrary number of covariates and the analytic establishment of its asymptotic properties. The second contribution is the efficient implementation of the estimator in practice. This is achieved by developing an automatic plug-in smoothing parameter selector which optimizes the estimator's performance in all coordinate directions. The traditional problem of vectorization of higher-order derivatives which lead to increasingly intractable matrix algebraic expressions is addressed here by introducing an alternative vectorization that exploits the analytic relationships between the functionals involved. This yields simpler, tractable and efficient in terms of computing time expressions which greatly facilitate the implementation of the rule in practice. The analytic study of the rule's rate of convergence shows that in contrast to the traditional cross validation approach, the proposed bandwidth selector is functional even for a large number of covariates. The benefits of all methodological advances are illustrated with the analysis of a motivating real-world dataset on credit risk.

Higher-order spatial autoregressive varying coefficient model: estimation and specification test P. 1258-1299

Tizheng Li ; Yuping Wang

Abstract

Conventional higher-order spatial autoregressive models assume that regression coefficients are constant over space, which is overly restrictive and unrealistic in applications. In this paper, we introduce higher-order spatial autoregressive varying coefficient model where regression coefficients are allowed to smoothly change over space, which enables us to simultaneously explore different types of spatial dependence and spatial heterogeneity of regression relationship. We propose a semi-parametric generalized method of moments estimation method for the proposed model and derive asymptotic properties of resulting estimators. Moreover, we propose a testing method to detect spatial heterogeneity of the regression relationship. Simulation studies show that the proposed estimation and testing methods perform quite well in finite samples. The Boston house price data are finally analyzed to demonstrate the proposed model and its estimation and testing methods.



The American Statistician, ISSN 0003-1305
Volume 79, number 1 (February 2025)

Abstract
An observational block design has I blocks matched for covariates and J individuals per block, but treatments were not randomly assigned to individuals within blocks, as would have been done in an experiment. Tightening an observational block design means selecting $J'<J$ individuals from each block, and possibly $I'\leq I$ blocks, to construct a new observational block design that, in some way, addresses unmeasured biases from nonrandom treatment assignment. Tightening must preserve covariate balance while altering the design to achieve some additional objective. An optimization algorithm is introduced that achieves this while maintaining the block structure by finely balancing covariates across blocks and through optimal subset matching. An example is considered in detail, both to motivate and illustrate the tightening of an observational block design. Two tightened designs are built from a study of light daily alcohol consumption and its possible effects on HDL cholesterol. One tightened design adjusts for an outcome tentatively presuming it was unaffected by the treatment. The second tightened design uses a differential effect to remove bias from an unobserved general disposition that promotes several treatments. An R package tightenBlock implements the method, contains the data, and in that package the help-file for the function tighten reproduces the example.

Abstract
Asymptotic goodness-of-fit methods in contingency table analysis can struggle with sparse data, especially in multi-way tables where it can be infeasible to meet sample size requirements for a robust application of distributional assumptions. However, algebraic statistics provides exact alternatives to these classical asymptotic methods that remain viable even with sparse data. We apply these methods to a context in psychometrics and education research that leads naturally to multi-way contingency tables: the analysis of differential item functioning (DIF). We explain concretely how to apply the exact methods of algebraic statistics to DIF analysis using the R package algstat, and we compare their performance to that of classical asymptotic methods.

Abstract
This article describes a simple and fast algorithm for generating correlation matrices (R) with a known average

correlation. The algorithm should be useful for researchers desiring plausible R matrices for substantive domains in which average correlations are known (at least approximately). The method is non-iterative and it can solve relatively large problems (e.g., generate a 500×500 R matrix) in less than a second on a personal computer. It also has didactic value for introducing students to the convex set of feasible R matrices of a fixed dimension. This Euclidean body is called an elliptope. The proposed method exploits the geometry of elliptopes to efficiently generate realistic R matrices with a desired average correlation coefficient. R code for implementing the algorithm (and for reproducing all of the results of this article) is reported in an online supplement.

The Best Time to Play the Lottery P. 30-39

Christopher M. Rump

Abstract

The best time to play the lottery is when the jackpot has rolled over several times and grown large, but not so large that you must share the prize if you win. We examine maximizing the expected value of a winning ticket as well as that in a random ticket. The derived optimality criteria depend on the prize elasticity of ticket demand. A regression analysis on data obtained from the Mega Millions® and Powerball® multi-state lotteries suggests ticket sales grow quadratically in the size of the advertised lump-sum cash jackpot prize. With quadratic growth, the best time to play is when ticket sales are 1.25–2.5 times the jackpot odds, currently about 300 M to one for these two lotteries. Since ticket sales are not known to ticket buyers, we invert the regression function to prescribe the best time to play in terms of the cash prize. It turns out that these lotteries offer a (pretax) fair wager with positive expected value in a surprisingly wide interval of jackpot prizes. That is a good time to play; the best time is in the neighborhood of the nearly 1 $B record cash jackpot awarded in these lotteries in recent years.

The R2D2 Prior for Generalized Linear Mixed Models P. 40-49

Eric Yanchenko, Howard D. Bondell & Brian J. Reich

Abstract

In Bayesian analysis, the selection of a prior distribution is typically done by considering each parameter in the model. While this can be convenient, in many scenarios it may be desirable to place a prior on a summary measure of the model instead. In this work, we propose a prior on the model fit, as measured by a Bayesian coefficient of determination (R2), which then induces a prior on the individual parameters. We achieve this by placing a beta prior on R^2 and then deriving the induced prior on the global variance parameter for generalized linear mixed models. We derive closed-form expressions in many scenarios and present several approximation strategies when an analytic form is not possible and/or to allow for easier computation. In these situations, we suggest approximating the prior by using a generalized beta prime distribution and provide a simple default prior construction scheme. This approach is quite flexible and can be easily implemented in standard Bayesian software. Lastly, we demonstrate the performance of the method on simulated and real-world data, where the method particularly shines in high-dimensional settings, as well as modeling random effects.

Sequential Monitoring Using the Second Generation P-Value with Type I Error Controlled by Monitoring Frequency P. 50-60

Jonathan J. Chipman, Robert A. Greevy Jr., Lindsay S. Mayberry & Jeffrey D. Blume

Abstract

The Second Generation P-Value (SGPV) measures the overlap between an estimated interval and a composite hypothesis of parameter values. We develop a sequential monitoring scheme of the SGPV (SeqSGPV) to connect study design intentions with end-of-study inference anchored on scientific relevance. We build upon Freedman's "Region of Equivalence" (ROE) in specifying scientifically meaningful hypotheses called Pre-specified Regions Indicating Scientific Merit (PRISM). We compare PRISM monitoring versus monitoring alternative ROE specifications. Error rates are controlled through the PRISM's indifference zone around the point null and monitoring frequency

strategies. Because the former is fixed due to scientific relevance, the latter is a targettable means for designing studies with desirable operating characters. An affirmation step to stopping rules improves frequency properties including the error rate, the risk of reversing conclusions under delayed outcomes, and bias.

Integrative Data Analysis Where Partial Covariates Have Complex Nonlinear Effects by Using Summary Information from an External Data P. 61-71

Jia Liang, Shuo Chen, Peter Kochunov, L. Elliot Hong & Chixiang Chen

Abstract

A full parametric and linear specification may be insufficient to capture complicated patterns in studies exploring complex features, such as those investigating age-related changes in brain functional abilities. Alternatively, a partially linear model (PLM) consisting of both parametric and nonparametric elements may have a better fit. This model has been widely applied in economics, environmental science, and biomedical studies. In this article, we introduce a novel statistical inference framework that equips PLM with high estimation efficiency by effectively synthesizing summary information from external data into the main analysis. Such an integrative scheme is versatile in assimilating various types of reduced models from the external study. The proposed method is shown to be theoretically valid and numerically convenient, and it ensures a high-efficiency gain compared to classic methods in PLM. Our method is further validated using two data applications by evaluating the risk factors of brain imaging measures and blood pressure.

High-Dimensional Propensity Score and Its Machine Learning Extensions in Residual Confounding Control P. 72-90

Mohammad Ehsanul Karim

Abstract

“The use of health care claims datasets often encounters criticism due to the pervasive issues of omitted variables and inaccuracies or mis-measurements in available confounders. Ultimately, the treatment effects estimated using such data sources may be subject to residual confounding. Digital electronic administrative records routinely collect a large volume of health-related information; and many of which are usually not considered in conventional pharmacoepidemiological studies. A high-dimensional propensity score (hdPS) algorithm was proposed that uses such information as surrogates or proxies for mismeasured and unobserved confounders in an effort to reduce residual confounding bias. Since then, many machine learning and semi-parametric extensions of this algorithm have been proposed to better exploit the wealth of high-dimensional proxy information. In this tutorial, we will (i) demonstrate logic, steps and implementation guidelines of hdPS using an open data source as an example (using reproducible R codes), (ii) familiarize readers with the key difference between propensity score versus hdPS, as well as the requisite sensitivity analyses, (iii) explain the rationale for using the machine learning and double robust extensions of hdPS, and (iv) discuss advantages, controversies, and hdPS reporting guidelines while writing amanuscript.

A Multi-Method Data Science Pipeline for Analyzing Police Service P. 91-101

Anna Haensch, Daanika Gordon, Karin Knudson & Justina Cheng

Abstract

Despite the fact that most police departments in the U.S. serve jurisdictions with fewer than 10,000 residents, policing practices in small towns are understudied. This is due in part to data limitations and technological barriers that exist in the small-town context. In this article we focus on one small town police department in New England with a history of misconduct, and develop a comprehensive data science pipeline that addresses the stages from design and collection to reporting. We present the reader with specific tools in the open-source Python ecosystem for replicating this pipeline. Once these data are processed, we perform two statistical analyses in an attempt to better understand the provisions of service by the small-town police department of focus. First, we perform ecological

inference to estimate the rate at which residents are placing calls for service. Second, we model wait times using a negative binomal regression model to account for overdispersion in the data. We discuss data and model limitations arising through the pipeline creation and analysis process.

Assessment and Continuous Improvement of an Undergraduate Data Science Program P. 102-121

Nicholas Clark, Christopher Morrell & Mike Powell

Abstract

In recent years, there has been an explosion in the growth of undergraduate statistics and data science programs across the US. Simultaneously, there has been clear guidance written on curriculum development for both data science (De Veaux et al.) and statistics (Carver et al.) programs. While this was occurring, ABET (now simply an acronym, but previously standing for the Accreditation Board for Engineering and Technology), in coordination with organizations such as the American Statistical Association, developed accreditation criteria for Data Science programs. In this article, we discuss our journey through ABET accreditation and discuss how adopting ABET processes for continuous improvement strengthens a program's assessment process. We share best practices for working across multiple departments to collect data not only on individual courses, but also on the program as a whole. While the framework presented was initially established to support ABET accreditation, we argue that a properly executed program assessment should occur regardless of whether or not an institution is seeking ABET accreditation for their data science program. Throughout this article, we also discuss the extent to which ABET requirements naturally fit within our program's existing goals, including an assessment of how ABET requirements align with major ideas in the field of data science education.

Distance Covariance, Independence, and Pairwise Differences P. 122-128

Jakob Raymaekers & Peter J. Rousseeuw

Abstract

Distance covariance (Székely, Rizzo, and Bakirov) is a fascinating recent notion, which is popular as a test for dependence of any type between random variables X and Y. This approach deserves to be touched upon in modern courses on mathematical statistics. It makes use of distances of the type $|X–X'|$ and $|Y–Y'|$, where (X',Y') is an independent copy of (X, Y). This raises natural questions about independence of variables like $X–X'$ and $Y–Y'$, about the connection between $\text{cov}(|X–X'|,|Y–Y'|)$ and the covariance between doubly centered distances, and about necessary and sufficient conditions for independence. We show some basic results and present a new and nontechnical counterexample to a common fallacy, which provides more insight. We also show some motivating examples involving bivariate distributions and contingency tables, which can be used as didactic material for introducing distance correlation.

A Review of Design of Experiments Courses Offered to Undergraduate Students at American Universities P. 129-139

Alan R. Vazquez & Xiaocong Xuan

Abstract

Design of Experiments (DoE) is a relevant class to undergraduate students in the sciences, because it teaches them how to plan, conduct, and analyze experiments. In the literature on DoE, there are several contributions to its pedagogy, such as easy-to-use class experiments, virtual experiments, and software to construct experimental designs. However, there are virtually no systematic evaluations of the actual DoE pedagogy. To address this issue, we build the first database of DoE courses offered to undergraduate students in the United States. The database has records on courses offered from 2019 to 2022 at the best universities in the US News Best National Universities ranking of 2022. Specifically, it has data on 18 general and content-specific features of 206 courses. To study the DoE pedagogy, we analyze the database using descriptive statistics and text mining. Based on our analysis, we

provide instructors with recommendations and teaching material to enhance their DoE courses. The database and material are included in the supplement of this article.

TOP : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1134-5764
Volume 32, number 1 (april 2024)

Abstract
Autonomous vehicles create new opportunities as well as new challenges to dynamic vehicle routing. The introduction of autonomous vehicles as information-collecting agents results in scenarios, where dynamic nodes are found by proximity. This paper presents a novel dynamic vehicle-routing problem variant with proximity-dependent nodes. Here, we introduced a novel variable, *detectability*, which determines whether a proximal dynamic node will be detected, based on the sight radius of the vehicle. The problem considered is motivated by autonomous weed-spraying vehicles in large agricultural operations. This work is generalisable to many other autonomous vehicle applications. The first step to crafting a solution approach for the problem is to decide *when* reoptimisation should be triggered. Two reoptimisation trigger strategies are considered—exogenous and endogenous. Computational experiments compared the strategies for both the classical dynamic vehicle routing problem as well as the introduced variant. Experiments used extensive standardised vehicle-routing problem benchmarks with varying degrees of dynamism and geographical node distributions. The results showed that for both the classical problem and the novel variant, an endogenous trigger strategy is better in most cases, while an exogenous trigger strategy is only suitable when both detectability and dynamism are low. Furthermore, the optimal level of detectability was shown to be dependent on the combination of trigger, degree of dynamism, and geographical node distribution, meaning practitioners may determine the required detectability based on the attributes of their specific problem.

Abstract
This paper presents a framework for the efficient management of renewable energies to charge a fleet of electric buses (eBuses). Our framework starts with the prediction of clean energy time windows, i.e., periods of time when the production of clean energy exceeds the demand of the country. Then, the optimization phase schedules charging events to reduce the use of non-clean energy to recharge eBuses while passengers are embarking or disembarking. The proposed framework is capable of overcoming the unstable and chaotic nature of wind power generation to operate the fleet without perturbing the quality of service. Our extensive empirical validation with real instances from Ireland suggests that our solutions can significantly reduce non-clean energy consumed on large data sets.

Abstract
Survival analysis is a family of statistical methods for analyzing event occurrence times. We adopt a mixed-integer

optimization approach to estimation of sparse Cox proportional hazards (PH) models for survival analysis. Specifically, we propose a high-performance cutting-plane algorithm based on a reformulation of our sparse estimation problem into a bilevel optimization problem. This algorithm solves the upper-level problem using cutting planes that are generated from the dual lower-level problem to approximate an upper-level nonlinear objective function. To solve the dual lower-level problem efficiently, we devise a quadratic approximation of the Fenchel conjugate of the loss function. We also develop a computationally efficient least-squares method for adjusting quadratic approximations to fit each dataset. Computational results demonstrate that our method outperforms regularized estimation methods in terms of accuracy for both prediction and subset selection especially for low-dimensional datasets. Moreover, our quadratic approximation of the Fenchel conjugate function accelerates the cutting-plane algorithm and maintains high generalization performance of sparse Cox PH models.

Berge equilibria and the equilibria of the altruistic game P. 83-105

A. Zapata ; A. M. Mármol ; L. Monroy

Abstract

Berge's notion of equilibrium represents a complementary alternative to the Nash equilibrium when modeling socioeconomic behavior and human interactions. While the notion of Nash equilibrium is based on self-interest, as players seek to maximize their own payoffs given the action of the other players, the idea behind Berge equilibrium is mutual support, as given the action of one of the players, all others select their actions looking for her best interest. However, because of the demanding conditions involved, the existence of Berge equilibria is rarely guaranteed. In this paper, we propose vector-valued normal-form games as a unified framework in which to study and extend the concept of Berge equilibrium. Based on the equilibria of the so-called altruistic game, we introduce new equilibrium concepts which constitute different relaxations of Berge's notion, although they still retain the underlying idea of mutual support. We establish the links between these new equilibria, Nash equilibrium, Berge equilibrium, and other related concepts already existing in the literature. Our approach has the advantage that it permits the incorporation of preference information to identify the equilibria which are consistent with different altruistic attitudes of the players.

Generating hydro unit commitment instances P. 106-136

Dimitri Thomopulos ; Wim van Ackooij ; Marc Stefanon

Abstract

Handling cascading reservoir systems is an important energy management optimization problem. The difficulty of these problems stems, in part, from the modeling of the hydro-production function. Data are not always easy to come by. To remedy this issue, this paper proposes and describes a realistic instance generator, building instances of varying difficulty. No specific data format for these problems is currently available in the literature. In addition, both notation and formulations tend to vary widely. Instances and case studies are poorly available or otherwise not applicable to different variants of these problems. The purpose of the generator is also a first attempt to develop a shared data format useful for different variants of hydro unit commitment problems and seeks to mediate the needs of different sectors interested in the optimization of hydropower plants.

On properties of the set of awards vectors for a claims problem P. 137–167

Miguel Ángel Mirás Calvo ; Lago Núñez Lugilde ; Estela Sánchez-Rodríguez

Abstract

We study the geometric structure of a particular type of nonempty convex polytopes that are the intersection of an n-rectangle with a hyperplane $x_1 + \ldots + x_n = E$, $E > 0$. This type of polytopes arise naturally when studying, for instance, the set of awards vectors for a claims problem, the core of the game associated with a bankruptcy problem, the core-cover set of a game, or the class of two-bound core games. We explore in detail the geometry of such a polytope and provide explicit expressions to compute its volume and its centroid. In particular, we describe a procedure to compute the average-of-awards rule for a claims problem directly from the parameters of the problem. We show that computing the average-of-awards rule is # P-complete.

TOP : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1134-5764
Volume 32, number 2 (july 2024)

Abstract
In the flow shop rescheduling literature, many papers consider unlimited buffer capacities between successive machines. In real fact, these capacities may be limited, or no store may exist. Thus, a blocking situation is inducted. Diverse types of blocking constraints are studied in the flow shop scheduling problems. However, in dynamic environments, only few papers deal with these kinds of constraints. The aim of this paper is to investigate a problem of rescheduling the jobs in a flowshop environment and mixed blocking as a constraint, considering simultaneously schedule efficiency and stability as a performance measure, and job arrival as a disruption. An iterative methodology based on the predictive–reactive strategy is implemented for dealing with this rescheduling problem. The problem has first been modeled as a Mixed Integer Linear Programing (MILP) model. Experimental results show that the MILP resolution is only possible for small-sized instances. Hence, inspired by NEH algorithm, we proposed four heuristics for solving large-sized instances of this problem. Eventually, we discussed the performance of the proposed heuristics for different blocking situations, both in terms of solution efficiency and resolution time.

Abstract
Wind power is a major source of green energy production. However, the energy generation of wind power is highly affected by uncertainty. Here, we consider the problem of designing the cable network that interconnects the turbines to the substation in wind farms, aiming to minimize both the infrastructure cost and the cost of the energy losses during the wind farm's lifetime. Nonetheless, the energy losses depend on wind direction and speed, which are rarely known with certainty in real situations. Hence, the design of the network should consider these losses as uncertain parameters. We assume that the exact probability distribution of these parameters is unknown but belongs to an ambiguity set and propose a distributionally robust two-stage mixed integer model. The model is solved using a decomposition algorithm. Three enhancements are proposed given the computational difficulty in solving real problem instances. Computational results are reported based on real data.

Abstract
Numerous operational constraints within both industry and service sectors mandate the concurrent scheduling of tasks. This need is particularly evident in the assembly of products within manufacturing processes. This paper concentrates on minimizing makespan in the two-stage assembly flow shop problem. Our contributions include the introduction of novel dominance rules, a proposal for a heuristic method, and the development of a branch and bound algorithm. Additionally, we conduct an empirical analysis of makespan distribution for small-size instances.

Through extensive experimentation, our study demonstrates the efficiency of the introduced dominance rules and the strong performance of the developed branch and bound algorithm.

Discussing some approaches to delta-shock modeling P. 245-262

Maxim Finkelstein ; Ji Hwan Cha

Abstract

We revisit the 'classical' delta-shock model and generalize it to the case of renewal processes of external shocks with arbitrary inter-arrival times and arbitrary distribution of the 'recovery' parameter delta. Our innovative approach is based on defining the renewal points for the model and deriving the corresponding integral equations for the survival probabilities of interest that describe the setting probabilistically. As examples, the cases of exponentially distributed and constant delta are analyzed. Furthermore, delta shock modeling for systems with protection and two shock processes is considered. The first process targets the defense system and can partially destroy it. In this case, the second process that targets the main, protected system can result in its failure. The damages of the defense system are recovered during the recovery time delta. As exact solutions of the discussed problems are rather cumbersome, we provide simple and easy approximate solutions that can be implemented in practice. These results are justified under the assumption of 'fast repair' when the recovery time delta is stochastically much smaller than the inter-arrival times of the shock processes. The corresponding numerical examples (with discussion) illustrate our findings.

An integrated production planning and inventory management problem for a perishable product: optimization and Monte Carlo simulation as a tool for planning in scenarios with uncertain demands P. 263-303

Jeferson Auto da Cruz ; Luiz Leduino de Salles-Neto ; Cleder Marcos Schenekemberg

Abstract

This paper addresses a novel problem inspired by a practical situation faced by a Brazilian dairy company, which presents the integration of production and a two-level storage inventory control. In the supply chain (SC) analyzed, multiple production plants, warehouses, and distribution centers are considered. The problem also involves a set of constraints related to logistical capabilities and conditions of material flow throughout the SC. We developed a mathematical model and a Monte Carlo Simulation routine that aims to assist managers in establishing a production and inventory management plan for a perishable product, considering a stochastic and non-stationary demand. A numerical illustration of the results obtained by the mathematical model and a case study are carried out on the consideration of a First-Expired, First-Out inventory management policy. Results demonstrate that the proposed method allows the analysis of safety stock and the achievement of desired service levels while promoting a reduced rate of food waste over the considered planning horizon.

An ALNS to optimize makespan subject to total completion time for no-wait flow shops with sequence-dependent setup times P. 304–322

Fernando Siqueira de Almeida ; Marcelo Seido Nagano

Abstract

In this article, we address the no-wait flow shop scheduling problem with sequence dependent setup times. The objective is to minimize makespan subject to an upper bound on total completion time. Although these performance measures and constraints have been extensively studied, they have never been considered together in this problem before. To solve the problem, we propose an adaptive large neighborhood search algorithm called *ALNSA*. Essentially, *ALNSA* improves an initial solution by dynamically selecting and executing a pair of destroy and repair methods based on their performance history. In addition to classic greedy and random methods used, we present two new mechanisms in which the greediness-randomness behavior is balanced. To evaluate performance, the proposed approach is compared with three heuristic methods—*GL*, *HH1* and *TOB*–developed for the most similar problems found in the literature. Computational experiments show that the proposed method outperforms state-of-the-art approaches in the literature for the no-wait flow shop scheduling problem with sequence dependent setup times and is therefore recommended to solve the problem.

Loss-optimal classification trees: a generalized framework and the logistic case P. 323–350

Tommaso Aldinucci ; Matteo Lapucci

Abstract

Classification trees are one of the most common models in interpretable machine learning. Although such models are usually built with greedy strategies, in recent years, thanks to remarkable advances in mixed-integer programming (MIP) solvers, several exact formulations of the learning problem have been developed. In this paper, we argue that some of the most relevant ones among these training models can be encapsulated within a general framework, whose instances are shaped by the specification of loss functions and regularizers. Next, we introduce a novel realization of this framework: specifically, we consider the logistic loss, handled in the MIP setting by a piece-wise linear approximation, and couple it with l_1 regularization terms. The resulting optimal logistic classification tree model numerically proves to be able to induce trees with enhanced interpretability properties and competitive generalization capabilities, compared to the state-of-the-art MIP-based approaches.

TOP : AN OFFICIAL JOURNAL OF THE SPANISH SOCIETY OF STATISTICS AND OPERATIONS RESEARCH, ISSN 1134-5764
Volume 32, number 3 (october 2024)

Abstract
Predicting the demographics of Twitter users has become a problem with a large interest in computational social sciences. However, the limited amount of public datasets with ground truth labels and the tremendous costs of hand-labeling make this task particularly challenging. Recently, programmatic weak supervision has emerged as a new framework to train classifiers on noisy data with minimal human labeling effort. In this paper, demographic prediction is framed for the first time as a programmatic weak supervision problem. A new three-step methodology for gender, age category, and location prediction is provided, which outperforms traditional programmatic weak supervision and is competitive with the state-of-the-art deep learning model. The study is performed in Flanders, a small Dutch-speaking European region, characterized by a limited number of user profiles and tweets. An evaluation conducted on an independent hand-labeled test set shows that the proposed methodology can be generalized to unseen users within the geographic area of interest.

Abstract
Among the most famous algorithms for solving classification problems are support vector machines (SVMs), which find a separating hyperplane for a set of labeled data points. In some applications, however, labels are only available for a subset of points. Furthermore, this subset can be non-representative, e.g., due to self-selection in a survey. Semi-supervised SVMs tackle the setting of labeled and unlabeled data and can often improve the reliability of the results. Moreover, additional information about the size of the classes can be available from undisclosed sources. We propose a mixed-integer quadratic optimization (MIQP) model that covers the setting of labeled and unlabeled data points as well as the overall number of points in each class. Since the MIQP's solution time rapidly grows as the number of variables increases, we introduce an iterative clustering approach to reduce the model's size. Moreover, we present an update rule for the required big-M values, prove the correctness of the iterative clustering method as well as derive tailored dimension-reduction and warm-starting techniques. Our numerical results show that our approach leads to a similar accuracy and precision than the MIQP formulation but at much lower computational cost. Thus, we can solve larger problems. With respect to the original SVM formulation, we observe that our approach has even better accuracy and precision for biased samples.

Abstract
Counterfactual explanations are increasingly used as an Explainable Artificial Intelligence (XAI) technique to provide stakeholders of

complex machine learning algorithms with explanations for data-driven decisions. The popularity of counterfactual explanations resulted in a boom in the algorithms generating them. However, not every algorithm creates uniform explanations for the same instance. Even though in some contexts multiple possible explanations are beneficial, there are circumstances where diversity amongst counterfactual explanations results in a potential disagreement problem among stakeholders. Ethical issues arise when for example, malicious agents use this diversity to fairwash an unfair machine learning model by hiding sensitive features. As legislators worldwide tend to start including the right to explanations for data-driven, high-stakes decisions in their policies, these ethical issues should be understood and addressed. Our literature review on the disagreement problem in XAI reveals that this problem has never been empirically assessed for counterfactual explanations. Therefore, in this work, we conduct a large-scale empirical analysis, on 40 data sets, using 12 explanation-generating methods, for two black-box models, yielding over 192,000 explanations. Our study finds alarmingly high disagreement levels between the methods tested. A malicious user is able to both exclude and include desired features when multiple counterfactual explanations are available. This disagreement seems to be driven mainly by the data set characteristics and the type of counterfactual algorithm. XAI centers on the transparency of algorithmic decision-making, but our analysis advocates for transparency about this self-proclaimed transparency.

Adapting support vector optimisation algorithms to textual gender classification P. 463-488

Javier Gomez ; Cesar Alfaro ; Raul Moreno

Abstract

In this paper, we focus on the problem of determining the gender of the person described in a biographical text. Since support vector machine classifiers are well suited for text classification tasks, we present a new stopping criterion for support vector optimisation algorithms tailored to this problem. This new approach exploits the geometric properties of the vector representation of such content. An experiment on a set of English and Spanish biographical articles retrieved from Wikipedia illustrates this approach and compares it to other machine learning classification algorithms. The proposed method allows real-time classification algorithm training. Moreover, these results confirm the advantage of leveraging additional gender information in strongly inflected languages, like Spanish, for this task.

Learning-assisted optimization for transmission switching P. 489–516

Salvador Pineda ; Juan Miguel Morales ; Asunción Jiménez-Cordero

Abstract

The design of new strategies that exploit methods from machine learning to facilitate the resolution of challenging and large-scale mathematical optimization problems has recently become an avenue of prolific and promising research. In this paper, we propose a novel learning procedure to assist in the solution of a well-known computationally difficult optimization problem in power systems: The Direct Current Optimal Transmission Switching (DC-OTS) problem. The DC-OTS problem consists in finding the configuration of the power network that results in the cheapest dispatch of the power generating units. With the increasing variability in the operating conditions of power grids, the DC-OTS problem has lately sparked renewed interest, because operational strategies that include topological network changes have proved to be effective and efficient in helping maintain the balance between generation and demand. The DC-OTS problem includes a set of binaries that determine the on/off status of the switchable transmission lines. Therefore, it takes the form of a mixed-integer program, which is NP-hard in general. In this paper, we propose an approach to tackle the DC-OTS problem that leverages known solutions to past instances of the problem to speed up the mixed-integer optimization of a new unseen model. Although our approach does not offer optimality guarantees, a series of numerical experiments run on a real-life power system dataset show that it features a very high success rate in identifying the optimal grid topology (especially when compared to alternative competing heuristics), while rendering remarkable speed-up factors.

Bolstering stochastic gradient descent with model building P. 517–536

Ş. İlker Birbil ; Özgür Martin ; Figen Öztoprak

Abstract

Stochastic gradient descent method and its variants constitute the core optimization algorithms that achieve good convergence rates for solving machine learning problems. These rates are obtained especially when these algorithms

are fine-tuned for the application at hand. Although this tuning process can require large computational costs, recent work has shown that these costs can be reduced by line search methods that iteratively adjust the step length. We propose an alternative approach to stochastic line search by using a new algorithm based on forward step model building. This model building step incorporates second-order information that allows adjusting not only the step length but also the search direction. Noting that deep learning model parameters come in groups (layers of tensors), our method builds its model and calculates a new step for each parameter group. This novel diagonalization approach makes the selected step lengths adaptive. We provide convergence rate analysis, and experimentally show that the proposed algorithm achieves faster convergence and better generalization in well-known test problems. More precisely, SMB requires less tuning, and shows comparable performance to other adaptive methods.

Gaining insight into crew rostering instances through ML-based sequential assignment P. 537–578

Philippe Racette ; Frédéric Quesnel ; François Soumis

Abstract

Crew scheduling is typically performed in two stages. First, solving the crew pairing problem generates sequences of flights called pairings. Then, the pairings are assigned to crew members to provide each person with a full schedule. A common way to do this is to solve an optimization problem called the crew rostering problem (CRP). However, before solving the CRP, the problem instance must be parameterized appropriately while taking different factors such as preassigned days off, crew training, sick leave, reserve duty, or unusual events into account. In this paper, we present a new method for the parameterization of CRP instances for pilots by scheduling planners. A machine learning-based sequential assignment procedure (*seqAsg*) whose arc weights are computed using a policy over state–action pairs for pilots is implemented to generate very fast solutions. We establish a relationship between the quality of the solutions generated by *seqAsg* and that of solutions produced by a state-of-the-art solver. Based on those results, we formulate recommendations for instance parameterization. Given that the *seqAsg* procedure takes only a few seconds to run, this allows scheduling workers to reparameterize crew rostering instances many times over the course of the planning process as needed.

Tuning parameters of deep neural network training algorithms pays off: a computational study P. 579–620

Corrado Coppola ; Lorenzo Papa ; Laura Palagi

Abstract

The paper aims to investigate the impact of the optimization algorithms on the training of deep neural networks with an eye to the interaction between the optimizer and the generalization performance. In particular, we aim to analyze the behavior of state-of-the-art optimization algorithms in relationship to their hyperparameters setting to detect robustness with respect to the choice of a certain starting point in ending on different local solutions. We conduct extensive computational experiments using nine open-source optimization algorithms to train deep Convolutional Neural Network architectures on an image multi-class classification task. Precisely, we consider several architectures by changing the number of layers and neurons per layer, to evaluate the impact of different width and depth structures on the computational optimization performance. We show that the optimizers often return different local solutions and highlight the strong correlation between the quality of the solution found and the generalization capability of the trained network. We also discuss the role of hyperparameters tuning and show how a tuned hyperparameters setting can be re-used for the same task on different problems achieving better efficiency and generalization performance than a default setting.