THE ANNALS OF APPLIED STATISTICS
AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Modelled approximations to the ideal filter with application to GDP and its components
THOMAS M. TRIMBUR AND TUCKER S. McELROY 627

Improving exoplanet detection power: Multivariate Gaussian process models for stellar activity
DAVID E. JONES, DAVID C. STENNING, ERIC B. FORD, ROBERT L. WOLPERT, THOMAS J. LOREDO,
CHRISTIAN GILBERTSON AND XAVIER DUMUSQUE 652

A Bayesian model of dose-response for cancer drug studies
WESLEY TANSEY, CHRISTOPHER TOSH AND DAVID M. BLEI 680

The assessment of replication success based on relative effect size
LEONHARD HELD, CHARLOTTE MICHELOUD AND SAMUEL PAWEL 706

Scalable change-point and anomaly detection in cross-correlated data with an application to condition monitoring
MARTIN TVETEN, IDRIS A. ECKLEY AND PAUL FEARNHEAD 721

Adaptive design for Gaussian process regression under censoring
JIALEI CHEN, SIMON MAK, V. ROSHAN JOSEPH AND CHUCK ZHANG 744

Composite mixture of log-linear models with application to psychiatric studies
EMANUELE ALIVERTI AND DAVID B. DUNSON 765

Inhomogeneous spatio-temporal point processes on linear networks for visitors’ stops data
NICOLETTA D’ANGELO, GIADA ADELFIO, ANTONINO ABBRUZZO AND JORGE MATEU 791

Batch-sequential design and heteroskedastic surrogate modeling for delta smelt conservation
BOYA ZHANG, ROBERT B. GRAMACY, LEAH R. JOHNSON, KENNETH A. ROSE AND ERIC SMITH 816

Intensity estimation on geometric networks with penalized splines
MARC SCHNEIBLE AND GÖRAN KAUERMANN 843

Sparse block signal detection and identification for shared cross-trait association analysis
JIANQIAO WANG, WANJIIE WANG AND HONGZHE LI 866

Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling with application to estimation of health insurance coverage
PAUL A. PARKER, SCOTT H. HOLAN AND RYAN JANICKI 887

Approximate Bayesian inference for analysis of spatiotemporal flood frequency data
ÁRNI V. JÖHANNesson, STEFAN SIEGERT, RAPHAËL HUSER, HAAKON BAKKA AND BIRGIR HRAFNIKELSSON 905

Permutation tests under a rotating sampling plan with clustered data
JIAHUA CHEN, YUKUN LIU, CARILYN G. TAYLOR AND JAMES V. IDEK 936

Inference for stochastic kinetic models from multiple data sources for joint estimation of infection dynamics from aggregate reports and virological data
OKSANA A. CHKREBTII, YURY E. GARCÍA, MARCOS A. CAPISTRÁN AND DANIEL E. NOYOLA 959

continued
Multistate capture–recapture models for irregularly sampled data
SINA MEWS, ROLAND LANGROCK, RUTH KING AND NICOLA QUICK 982
Bayesian inverse reinforcement learning for collective animal movement
TORYN L. J. SCHAFER, CHRISTOPHER K. WIKLE AND MEVIN B. HOOTEN 999
A flexible sensitivity analysis approach for unmeasured confounding with multiple
treatments and a binary outcome with application to SEER-Medicare lung cancer data
LIANGYUAN HU, JUNGANG ZOU, CHENYANG GU, JIAYI JI,
MICHAEL LOPEZ AND MINAL KALE 1014
Robust Bayesian inference for Big Data: Combining sensor-based records with traditional
survey data ........................................... ALI RAFEI, CAROL A. C. FLANNAGAN,
BRADY T. WEST AND MICHAEL R. ELLIOTT 1038
A sparse negative binomial classifier with covariate adjustment for RNA-seq data
TANBIN RAHMAN, HSIN-EN HUANG, YUHIA LI, AN-SHUN TAI,
WEN-PING HSEIH, COLLEEN A. MCCMUNG AND GEORGE TSENG 1071
Kernel machine and distributed lag models for assessing windows of susceptibility to
environmental mixtures in children’s health studies .................. ANDER WILSON,
HSIAO-HSIEH LEON HSU, YUEH-HSIU MATHILDA CHIU, ROBERT O. WRIGHT,
ROSALIND J. WRIGHT AND BRENT A. COULL 1090
Detecting heterogeneous treatment effects with instrumental variables and application to
the Oregon health insurance experiment .................... MICHAEL JOHNSON,
JIONGYI CAO AND HYUNSEUNG KANG 1111
Statistical shape analysis of brain arterial networks (BAN)
XIAOYANG GUO, ADITI BASU BAL, TOM NEEDHAM AND ANU SRIVASTAVA 1130
Spatiotemporal-textual point processes for crime linkage detection
SHIXIANG ZHU AND YAO XIE 1151
Markov-modulated Hawkes processes for modeling sporadic and bursty event occurrences
in social interactions ..................................... JING WU, OWEN G. WARD,
JAMES CURLEY AND TIAN ZHENG 1171
Conditional functional clustering for longitudinal data with heterogeneous nonlinear
patterns . . . . . TIANHAO WANG, LEI YU, SUE E. LEURGANS, ROBERT S. WILSON,
DAVID A. BENNETT AND PATRICIA A. BOYLE 1191
Impact evaluation of the LAPD community safety partnership . . . . . SYDNEY KAHMANN,
ERIN HARTMAN, JORJA LEAP AND P. JEFFREY BRANTINGHAM 1215
Higher criticism for discriminating word-frequency tables and authorship attribution
ALON KIPNIS 1236
INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

President: Krzysztof Burdzy, Department of Mathematics, University of Washington, Seattle, Washington 98195-4350, USA

President-Elect: Peter Bühlmann, Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland

Past President: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, USA

IMS PUBLICATIONS

The Annals of Statistics. Editors: Enno Mammen, Institute for Applied Mathematics, Heidelberg University, 69120 Heidelberg, Germany. Lan Wang, Miami Herbert Business School, University of Miami, Coral Gables, FL 33124, USA

The Annals of Applied Statistics. Editor-In-Chief: Ji Zhu, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

The Annals of Probability. Editors: Alice Guionnet, Unité de Mathématiques Pures et Appliquées, ENS de Lyon, Lyon, France. Christophe Garban, Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France

The Annals of Applied Probability. Editors: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Qi-Man Shao, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, P.R. China

Statistical Science. Editor: Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

The IMS Bulletin. Editor: Tati Howell, bulletin@imstat.org

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 16, Number 2, June 2022. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to The Annals of Applied Statistics, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Copyright © 2022 by the Institute of Mathematical Statistics
Printed in the United States of America
MODELLED APPROXIMATIONS TO THE IDEAL FILTER WITH APPLICATION TO GDP AND ITS COMPONENTS\textsuperscript{1}

BY THOMAS M. TRIMBUR\textsuperscript{a} AND TUCKER S. MCELROY\textsuperscript{b}

Center for Statistical Research and Methodology, U.S. Census Bureau, \textsuperscript{a}Thomas.Trimbur@census.gov, \textsuperscript{b}tucker.s.mcelroy@census.gov

This paper examines cyclical fluctuations in a comprehensive statistical application, focusing on U.S. macroeconomic indicators related to real gross domestic product (real GDP). While GDP is generally viewed as the most widespread measure of economic activity available, our dataset also encompasses the primary GDP components, such as investment, together with leading (and regularly analyzed) subcomponent series, like residential and inventory investment. Analysis of the cycles in these major sectors provides a more informative perspective on the macroeconomic state and may improve a researcher’s ability to understand and forecast cyclical movements and growth in GDP. Adaptive time series modelling is used for each time series to derive the preferred band-pass filter for computing the optimal cycle. This contrasts with the rigid use of the ideal filter, whose gain function is perfectly sharp. Regarding the ideal filter, we provide an improved implementation compared to current practice. Thus, a set of approximating filters is derived that allow for a more attractive gain profile, a better match to the targeted passband, and a direct statistical way to extract signals near the sample endpoints. Our application study demonstrates that the commonly used ideal filter can perform quite poorly on a routine basis and lead to incorrect conclusions about even the most basic questions about empirical cyclical properties. The amplitude of filtered economic activity can have major distortions and become expanded or diminished (depending on the GDP component under consideration), and many essential divergences in path may occur and affect key signals, such as expansion or contraction in growth. Statistical measures of model performance very strongly favor the adaptive parameter approach. Our statistical analysis reveals diverse dynamic behavior among the series; such results may yield worthwhile insights for output sector analysts and, even for those primarily focused on GDP, may lead to possible modelling improvements by using the finer information content in the GDP-component dynamics.

REFERENCES


\textit{Key words and phrases.} Business cycles, band-pass filter, ideal filter, signal extraction, stochastic cycles, unobserved components.


IMPROVING EXOPLANET DETECTION POWER: MULTIVARIATE GAUSSIAN PROCESS MODELS FOR STELLAR ACTIVITY

BY DAVID E. JONES1,a, DAVID C. STENNING2,b, ERIC B. FORD3,4,5,6,c, ROBERT L. WOLPERT7,e, THOMAS J. LOREDO8,f, CHRISTIAN GILBERTSON3,d AND XAVIER DUMUSQUE9,g

1Department of Statistics, Texas A&M University, david.jones@tamu.edu
2Department of Statistics and Actuarial Science, Simon Fraser University, dstenning@sfu.ca
3Department of Astronomy and Astrophysics, Pennsylvania State University, ebf11@psu.edu, cjg66@psu.edu
4Center for Exoplanets & Habitable Worlds, Pennsylvania State University
5Center for Astrostatistics, Pennsylvania State University
6Institute for Computational & Data Sciences, Pennsylvania State University
7Department of Statistical Science, Duke University, rlw@duke.edu
8Cornell Center for Astrophysics and Planetary Science, Cornell University, loredo@astro.cornell.edu
9Observatoire Astronomique de l’Université de Genève, Xavier.Dumusque@unige.ch

The radial velocity method is one of the most successful techniques for detecting exoplanets. It works by detecting the velocity of a host star, induced by the gravitational effect of an orbiting planet, specifically, the velocity along our line of sight which is called the radial velocity of the star. Low-mass planets typically cause their host star to move with radial velocities of 1 m/s or less. By analyzing a time series of stellar spectra from a host star, modern astronomical instruments can, in theory, detect such planets. However, in practice, intrinsic stellar variability (e.g., star spots, convective motion, pulsations) affects the spectra and often mimics a radial velocity signal. This signal contamination makes it difficult to reliably detect low-mass planets. A principled approach to recovering planet radial velocity signals in the presence of stellar activity was proposed by Rajpaul et al. (Mon. Not. R. Astron. Soc. 452 (2015) 2269–2291). It uses a multivariate Gaussian process model to jointly capture time series of the apparent radial velocity and multiple indicators of stellar activity. We build on this work in two ways: (i) we propose using dimension reduction techniques to construct new high-information stellar activity indicators; and (ii) we extend the Rajpaul et al. (Mon. Not. R. Astron. Soc. 452 (2015) 2269–2291) model to a larger class of models and use a power-based model comparison procedure to select the best model. Despite significant interest in exoplanets, previous efforts have not performed large-scale stellar activity model selection or attempted to evaluate models based on planet detection power. In the case of main sequence G2V stars, we find that our method substantially improves planet detection power, compared to previous state-of-the-art approaches.

REFERENCES


Key words and phrases. Exoplanets, radial velocity method, stellar activity, Gaussian process, time series, dimension reduction, model selection.


A BAYESIAN MODEL OF DOSE-RESPONSE FOR CANCER DRUG STUDIES

BY WESLEY TANSEY¹,a, CHRISTOPHER TOSH²,b AND DAVID M. BLEI²,c

¹Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, a tanseyw@mskcc.org
²Data Science Institute, Columbia University, bct2915@columbia.edu, c david.blei@columbia.edu

Exploratory cancer drug studies test multiple tumor cell lines against multiple candidate drugs. The goal in each paired (cell line, drug) experiment is to map out the dose-response curve of the cell line as the dose level of the drug increases. We propose Bayesian tensor filtering (BTF), a hierarchical Bayesian model for dose-response modeling in multisample, multitreatment cancer drug studies. BTF uses low-dimensional embeddings to share statistical strength between similar drugs and similar cell lines. Structured shrinkage priors in BTF encourage smoothness in the dose-response curves while remaining adaptive to sharp jumps when the data call for it. We focus on a pair of cancer drug studies exhibiting a particular pathology in their experimental design, leading us to a nonconjugate monotone mixture-of-gammas likelihood. To perform posterior inference, we develop a variant of the elliptical slice sampling algorithm for sampling from linearly-constrained multivariate normal priors with nonconjugate likelihoods. In benchmarks, BTF outperforms state-of-the-art methods for covariance regression and dynamic Poisson matrix factorization. On the two cancer drug studies, BTF outperforms the current standard approach in biology and reveals potential new biomarkers of drug sensitivity in cancer. Code is available at https://github.com/tansey/functionalmf.

REFERENCES


Key words and phrases. Dose-response, matrix factorization, trend filtering, slice sampling, constrained inference.


THE ASSESSMENT OF REPPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

BY LEONHARD HELD\textsuperscript{a}, CHARLOTTE MICHELOUD\textsuperscript{b} AND SAMUEL PAWEL\textsuperscript{c}

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich, \textsuperscript{a}leonhard.held@uzh.ch, \textsuperscript{b}charlotte.micheloud@uzh.ch, \textsuperscript{c}samuel.pawel@uzh.ch

Replication studies are increasingly conducted in order to confirm original findings. However, there is no established standard how to assess replication success, and, in practice, many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that, for borderline significant original studies, replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the type-I error rate if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate, compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

REFERENCES


\textit{Key words and phrases.} Power, replication studies, sceptical \textit{p}-value, shrinkage, two-trials rule, type-I error rate.

SCALABLE CHANGE-POINT AND ANOMALY DETECTION IN CROSS-CORRELATED DATA WITH AN APPLICATION TO CONDITION MONITORING

BY MARTIN TVETEN\textsuperscript{1,a}, IDRIS A. ECKLEY\textsuperscript{2,b} AND PAUL FEARNHEAD\textsuperscript{2,c}

\textsuperscript{1}Department of Mathematics, University of Oslo, \textsuperscript{a}tveten@nr.no
\textsuperscript{2}Mathematics and Statistics, Lancaster University, \textsuperscript{b}i.eckley@lancaster.ac.uk, \textsuperscript{c}p.fearnhead@lancaster.ac.uk

Motivated by a condition monitoring application arising from subsea engineering, we derive a novel, scalable approach to detecting anomalous mean structure in a subset of correlated multivariate time series. Given the need to analyse such series efficiently, we explore a computationally efficient approximation of the maximum likelihood solution to the resulting modelling framework and develop a new dynamic programming algorithm for solving the resulting binary quadratic programme when the precision matrix of the time series at any given time point is banded. Through a comprehensive simulation study we show that the resulting methods perform favorably compared to competing methods, both in the anomaly and change detection settings, even when the sparsity structure of the precision matrix estimate is misspecified. We also demonstrate its ability to correctly detect faulty time periods of a pump within the motivating application.

REFERENCES


Key words and phrases. Anomaly, binary quadratic programme, change points, cross-correlation, outliers.


ADAPTIVE DESIGN FOR GAUSSIAN PROCESS REGRESSION UNDER CENSORING

BY JIALEI CHEN¹,a, SIMON MAK²,d, V. ROSHAN JOSEPH¹,b AND CHUCK ZHANG¹,c

¹H. Milton Stewart School of Industrial & Systems Engineering Georgia Institute of Technology, ²Department of Statistical Science, Duke University

Adaptive sampling, censored experiments, experimental design, kriging, multifidelity modeling.


COMPOSITE MIXTURE OF LOG-LINEAR MODELS WITH APPLICATION TO PSYCHIATRIC STUDIES

BY EMANUELE ALIVERTI1,a AND DAVID B. DUNSON2,b

1Department of Economics, University Ca’ Foscari Venezia, aemanuele.aliverti@unive.it
2Department of Statistical Sciences, Duke University, bdunson@duke.edu

Psychiatric studies of suicide provide fundamental insights on the evolution of severe psychopathologies, and contribute to the development of early treatment interventions. Our focus is on modelling different traits of psychosis and their interconnections, focusing on a case study on suicide attempt survivors. Such aspects are recorded via multivariate categorical data, involving a large numbers of items for multiple subjects. Current methods for multivariate categorical data—such as penalized log-linear models and latent structure analysis—are either limited to low-dimensional settings or include parameters with difficult interpretation. Motivated by this application, this article proposes a new class of approaches, which we refer to as Mixture of Log Linear models (MILLS). Combining latent class analysis and log-linear models, MILLS defines a novel Bayesian approach to model complex multivariate categorical data with flexibility and interpretability, providing interesting insights on the relationship between psychotic diseases and psychological aspects in suicide attempt survivors.

REFERENCES


Key words and phrases. Psychiatric profiles, Bayesian modelling, categorical data, contingency table, log-linear models, mixture model.


INHOMOGENEOUS SPATIO-TEMPORAL POINT PROCESSES ON LINEAR NETWORKS FOR VISITORS’ STOPS DATA

BY NICOLETTA D’ANGELO\(^1,a\), GIADA ADELFI\(^1,b\), ANTONINO ABRUZZO\(^1,c\) AND JORGE MATEU\(^2,d\)

\(^1\)Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, \(^a\)nicoletta.dangelo@unipa.it, \(^b\)giada.adelfio@unipa.it, \(^c\)antonino.abbruzzo@unipa.it
\(^2\)Department of Mathematics, University Jaume I, \(^d\)mateu@uji.es

We analyse the spatio-temporal distribution of visitors’ stops by touristic attractions in Palermo (Italy), using theory of stochastic point processes living on linear networks. We first propose an inhomogeneous Poisson point process model with a separable parametric spatio-temporal first-order intensity. We account for the spatial interaction among points on the given network, fitting a Gibbs point process model with mixed effects for the purely spatial component. This allows us to study first-order and second-order properties of the point pattern, accounting both for the spatio-temporal clustering and interaction and for the spatio-temporal scale at which they operate. Due to the strong degree of clustering in the data, we then formulate a more complex model, fitting a spatio-temporal log-Gaussian Cox process to the point process on the linear network, addressing the problem of the choice of the most appropriate distance metric.

REFERENCES


Key words and phrases. Gibbs point processes, global positioning system, intensity estimation, linear networks, log-Gaussian Cox processes, spatio-temporal point processes.


Batch-Sequential Design and Heteroskedastic Surrogate Modeling for Delta Smelt Conservation

By Boya Zhang\textsuperscript{1,a}, Robert B. Gramacy\textsuperscript{1,b}, Leah R. Johnson\textsuperscript{1,c}, Kenneth A. Rose\textsuperscript{2,e} and Eric Smith\textsuperscript{1,d}

\textsuperscript{1}Department of Statistics, Virginia Tech, \texttt{boya66@vt.edu}, \textsuperscript{b}rbg@vt.edu, \textsuperscript{c}lrjohn@vt.edu, \textsuperscript{d}epsmith@vt.edu
\textsuperscript{2}University of Maryland Center for Environmental Science, Horn Point Laboratory, \texttt{krose@umces.edu}

Delta smelt is an endangered fish species in the San Francisco estuary that have shown an overall population decline over the past 30 years. Researchers have developed a stochastic, agent-based simulator to virtualize the system with the goal of understanding the relative contribution of natural and anthropogenic factors that might play a role in their decline. However, the input configuration space is high dimensional, running the simulator is time-consuming, and its noisy outputs change nonlinearly in both mean and variance. Getting enough runs to effectively learn input–output dynamics requires both a nimble modeling strategy and parallel evaluation. Recent advances in heteroskedastic Gaussian process (HetGP) surrogate modeling helps, but little is known about how to appropriately plan experiments for highly distributed simulation. We propose a batch sequential design scheme, generalizing one-at-a-time variance-based active learning for HetGP, as a means of keeping multicore cluster nodes fully engaged with runs. Our acquisition strategy is carefully engineered to favor selection of replicates which boost statistical and computational efficiency when training surrogates to isolate signal from noise. Design and modeling are illustrated on a range of toy examples before embarking on a large-scale smelt simulation campaign and downstream high-fidelity input sensitivity analysis.

REFERENCES


Key words and phrases. Gaussian process surrogate modeling, agent-based model, active learning, input-dependent noise, replication, sensitivity analysis.


INTENSITY ESTIMATION ON GEOMETRIC NETWORKS WITH PENALIZED SPLINES

BY MARC SCHNEBLEa AND GÖRAN KAUERMANNb

Department of Statistics, Ludwig-Maximilians-Universität München, a marc.schneble@stat.uni-muenchen.de, b goeran.kauermann@stat.uni-muenchen.de

In the past decades the growing amount of network data lead to many novel statistical models. In this paper we consider so-called geometric networks. Typical examples are road networks or other infrastructure networks. Nevertheless, the neurons or the blood vessels in a human body can also be interpreted as a geometric network embedded in a three-dimensional space. A network-specific metric, rather than the Euclidean metric, is usually used in all these applications, making the analyses of network data challenging. We consider network-based point processes, and our task is to estimate the intensity (or density) of the process which allows us to detect high- and low-intensity regions of the underlying stochastic processes. Available routines that tackle this problem are commonly based on kernel smoothing methods. This paper uses penalized spline smoothing and extends this toward smooth intensity estimation on geometric networks. Furthermore, our approach easily allows incorporating covariates, enabling us to respect the network geometry in a regression model framework. Several data examples and a simulation study show that penalized spline-based intensity estimation on geometric networks is a numerically stable and efficient tool. Furthermore, it also allows estimating linear and smooth covariate effects, distinguishing our approach from already existing methodologies.

REFERENCES


Key words and phrases. Intensity estimation of stochastic point processes, generalized additive models, geometric networks, penalized splines, poisson regression with offset, spatstat package.
MR2206355

https://doi.org/10.1111/biom.12666

SPARSE BLOCK SIGNAL DETECTION AND IDENTIFICATION FOR SHARED CROSS-TRAIT ASSOCIATION ANALYSIS

BY JIANQIAO WANG\textsuperscript{1,a}, WANJIE WANG\textsuperscript{2,c} AND HONGZHE LI\textsuperscript{1,b}

\textsuperscript{1}Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, \textsuperscript{a}wang jq@upenn.edu, \textsuperscript{b}hongzhe@upenn.edu

\textsuperscript{2}Department of Statistics and Applied Probability, National University of Singapore, \textsuperscript{c}staww@nus.edu.sg

Genome-wide association studies (GWAS) have identified thousands of single nucleotide polymorphisms (SNPs) that are associated with complex traits. GWAS data allows us to investigate the shared genetic etiologies among different traits. However, linkage disequilibrium (LD) between the SNPs complicates the detection and identification of shared genetic effects. In this paper we model the LD by dividing the genome into LD blocks and linking the genetic variants within a block to a possible latent causal variant. An eigenvector-projected score statistic that leverages the set of variants in LD and a maxtype test statistic (Max-block) are proposed to detect the existence of cross-trait genetic association. The Max-block is easy to calculate and is shown to control the genome-wide error rate. After the detection a stepwise procedure is proposed to identify the significant blocks that explain the genetic sharing between two traits. Simulation experiments show that Max-block is more powerful than standard approaches in the sparse settings and is robust to different signal strengths or levels of sparsity. The method is applied to study shared cross-trait associations in 10 pediatric autoimmune diseases and identified several regions that explain the genetic sharing between juvenile idiopathic arthritis (JIA) and ulcerative colitis (UC) and between UC and Crohn’s disease (CD). In addition, our analysis also indicates the genetic sharing in the MHC region among systemic lupus (SLE), celiac disease (CEL) and common variable immunodeficiency (CVID). Results from real data and simulation studies show that Max-block provides an important alternative to commonly used genetic correlation estimation in understanding genetic correlation among complex diseases.

REFERENCES


\textit{Key words and phrases}. Local dependency, pleiotropy, principal component analysis, sparse simultaneous signal.
COMPUTATIONALLY EFFICIENT BAYESIAN UNIT-LEVEL MODELS FOR NON-GAUSSIAN DATA UNDER INFORMATIVE SAMPLING WITH APPLICATION TO ESTIMATION OF HEALTH INSURANCE COVERAGE

BY PAUL A. PARKER¹,a, SCOTT H. HOLAN¹,2,b AND RYAN JANICKI³,c

¹Department of Statistics, University of Missouri, apaulparker@mail.missouri.edu, bholans@missouri.edu
²Office of the Associate Director for Research and Methodology, U.S. Census Bureau
³Center for Statistical Research and Methodology, U.S. Census Bureau, cryan.janicki@census.gov

Statistical estimates from survey samples have traditionally been obtained via design-based estimators. In many cases these estimators tend to work well for quantities, such as population totals or means, but can fall short as sample sizes become small. In today’s “information age,” there is a strong demand for more granular estimates. To meet this demand, using a Bayesian pseudolikelihood, we propose a computationally efficient unit-level modeling approach for non-Gaussian data collected under informative sampling designs. Specifically, we focus on binary and multinomial data. Our approach is both multivariate and multiscale, incorporating spatial dependence at the area level. We illustrate our approach through an empirical simulation study and through a motivating application to health insurance estimates, using the American Community Survey.

REFERENCES


Key words and phrases. Bayesian analysis, informative sampling, Pólya Gamma, pseudolikelihood, small area estimation, Small Area Health Insurance Estimates (SAHIE) Program.
Approximate Bayesian Inference for Analysis of Spatiotemporal Flood Frequency Data

By Árni V. Jóhannesson¹,a, Stefan Siegert²,c, Raphaël Huser³,d, Haakon Bakka⁴,e, and Birgir Hrafnkelsson¹,b

¹Department of Mathematics, Faculty of Physical Sciences, School of Engineering and Natural Sciences, University of Iceland, aavj2@hi.is, bbirgirhr@hi.is
²Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, cs.siegert@exeter.ac.uk
³Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), draphael.huser@kaust.edu.sa
⁴Department of Mathematics, University of Oslo, ebakka@r-inla.org

Extreme floods cause casualties and widespread damage to property and vital civil infrastructure. Predictions of extreme floods, within gauged and ungauged catchments, is crucial to mitigate these disasters. In this paper a Bayesian framework is proposed for predicting extreme floods, using the generalized extreme-value (GEV) distribution. A major methodological challenge is to find a suitable parametrization for the GEV distribution when multiple covariates and/or latent spatial effects are involved and a time trend is present. Other challenges involve balancing model complexity and parsimony, using an appropriate model selection procedure and making inference based on a reliable and computationally efficient approach. We here propose a latent Gaussian modeling framework with a novel multivariate link function designed to separate the interpretation of the parameters at the latent level and to avoid unreasonable estimates of the shape and time trend parameters. Structured additive regression models, which include catchment descriptors as covariates and spatially correlated model components, are proposed for the four parameters at the latent level. To achieve computational efficiency with large datasets and richly parametrized models, we exploit a highly accurate and fast approximate Bayesian inference approach which can also be used to efficiently select models separately for each of the four regression models at the latent level. We applied our proposed methodology to annual peak river flow data from 554 catchments across the United Kingdom. The framework performed well in terms of flood predictions for both ungauged catchments and future observations at gauged catchments. The results show that the spatial model components for the transformed location and scale parameters as well as the time trend are all important, and none of these should be ignored. Posterior estimates of the time trend parameters correspond to an average increase of about 1.5% per decade with range 0.1% to 2.8% and reveal a spatial structure across the United Kingdom. When the interest lies in estimating return levels for spatial aggregates, we further develop a novel copula-based postprocessing approach of posterior predictive samples in order to mitigate the effect of the conditional independence assumption at the data level, and we demonstrate that our approach indeed provides accurate results.

References


Key words and phrases. Approximate Bayesian inference, flood frequency analysis, latent Gaussian model, Max-and-Smooth, multivariate link function, spatiotemporal extremes.


PERMUTATION TESTS UNDER A ROTATING SAMPLING PLAN WITH CLUSTERED DATA

BY JIAHUA CHEN\textsuperscript{1,a}, YUKUN LIU\textsuperscript{2,d}, CARILYN G. TAYLOR\textsuperscript{1,b} AND JAMES V. ZIDEK\textsuperscript{1,c}

\textsuperscript{1}Department of Statistics, University of British Columbia, \textsuperscript{a}jhchen@stat.ubc.ca, \textsuperscript{b}cgtaylor@stat.ubc.ca, \textsuperscript{c}jim@stat.ubc.ca
\textsuperscript{2}School of Statistics, East China Normal University, \textsuperscript{d}ykliu@sfs.ecnu.edu.cn

The distribution of lumber strength of any grade may evolve, for example, due to climate change, forest fire, changes in processing methods, and other factors. So, in North America the forest products industry monitors the evolution of their means, percentiles, or other parameters to ensure the wood products meet the industrial standard. For administrative convenience and informativeness, one may adopt a rotating sampling plan by sampling 36 mills in the initial occasion and having six of them replaced in each successive occasion for the next five occasions. The strength data on a specified number, commonly 10 pieces of lumbers from each sampled mills, are obtained. Under such rotating plans the observations on pieces from the same mill are correlated, and the observations on samples from the same mill taken on different occasions are also correlated. Ignoring these correlations may lead to invalid inference procedures. Yet accommodating a cluster structure in parametric models is difficult and entails a high level of misspecification risk. In this paper we explore symmetry in the clustered data collected via a rotating sampling plan to develop a permutation scheme for testing various hypotheses of interest. We also introduce a semiparametric density ratio model to link the distributions of the response variable over time. The combination retains the validity of the inference methods while extracting maximum information from the sampling plan. A simulation study indicates that the proposed permutation tests firmly control the type I error whether or not the data are clustered. The use of the density ratio model improves the power of the tests. We also apply the proposed tests to data from the motivating application. The proposed permutation tests effectively address many real-world issues with trust worth inference conclusions.

REFERENCES


Key words and phrases. Density ratio model, empirical likelihood, exchangeability, multiple sample, percentila, rank test.
INFERENCE FOR STOCHASTIC KINETIC MODELS FROM MULTIPLE DATA SOURCES FOR JOINT ESTIMATION OF INFECTION DYNAMICS FROM AGGREGATE REPORTS AND VIROLOGICAL DATA

BY OKSANA A. CHKREBTII1,a, YURY E. GARCÍA2,b, MARCOS A. CAPISTRÁN2,c AND DANIEL E. NOYOLA3,d

1Department of Statistics, The Ohio State University, aoksana@stat.osu.edu
2Área de Matemáticas Básicas, Centro de Investigación en Matemáticas, byury@cimat.mx, cmarcos@cimat.mx
3Department of Microbiology, Faculty of Medicine, Universidad Autónoma de San Luis Potosí, dnoyola@uaslp.mx

Before the current pandemic, influenza and respiratory syncytial virus (RSV) were the leading etiological agents of seasonal acute respiratory infections (ARI) around the world. In this setting, medical doctors typically based the diagnosis of ARI on patients’ symptoms alone and did not routinely conduct virological tests necessary to identify individual viruses, limiting the ability to study the interaction between multiple pathogens and to make public health recommendations. We consider a stochastic kinetic model (SKM) for two interacting ARI pathogens circulating in a large population and an empirically-motivated background process for infections with other pathogens causing similar symptoms. An extended marginal sampling approach, based on the linear noise approximation to the SKM, integrates multiple data sources and additional model components. We infer the parameters defining the pathogens’ dynamics and interaction within a Bayesian model and explore the posterior trajectories of infections for each illness based on aggregate infection reports from six epidemic seasons collected by the state health department and a subset of virological tests from a sentinel program at a general hospital in San Luis Potosí, México. We interpret the results and make recommendations for future data collection strategies.

REFERENCES


Key words and phrases. Stochastic kinetic models, acute respiratory disease, Bayesian modeling, linear noise approximation, influenza, RSV.


MULTISTATE CAPTURE–RECAPTURE MODELS FOR IRREGULARLY SAMPLED DATA

BY SINA Mews1,a, ROLAND LANGROCK1,b, RUTH KING2,c AND NICOLA QUICK3,d

1Department of Business Administration and Economics, Bielefeld University, a sina.mews@uni-bielefeld.de, b roland.langrock@uni-bielefeld.de
2School of Mathematics, University of Edinburgh, c Ruth.King@ed.ac.uk
3School of Biology, University of St Andrews, d nicola.quick@duke.edu

Multistate capture-recapture data comprise individual-specific sighting histories, together with information on individuals’ states related, for example, to breeding status, infection level, or geographical location. Such data are often analysed using the Arnason–Schwarz model, where transitions between states are modelled using a discrete-time Markov chain, making the model most easily applicable to regular time series. When time intervals between capture occasions are not of equal length, more complex time-dependent constructions may be required, increasing the number of parameters to estimate, decreasing interpretability, and potentially leading to reduced precision. Here we develop a multi-state model based on a state process operating in continuous time, which can be regarded as an analogue of the discrete-time Arnason–Schwarz model for irregularly sampled data. Statistical inference is carried out by regarding the capture-recapture data as realisations from a continuous-time hidden Markov model, which allows the associated efficient algorithms to be used for maximum likelihood estimation and state decoding. To illustrate the feasibility of the modelling framework, we use a long-term survey of bottlenose dolphins where capture occasions are not regularly spaced through time. Here, we are particularly interested in seasonal effects on the movement rates of the dolphins along the Scottish east coast. The results reveal seasonal movement patterns between two core areas of their range, providing information that will inform conservation management.

REFERENCES


Key words and phrases. Arnason–Schwarz model, continuous-time Markov chain, hidden Markov model, maximum likelihood.


Agent-based methods allow for defining simple rules that generate complex group behaviors. The governing rules of such models are typically set a priori, and parameters are tuned from observed behavior trajectories. Instead of making simplifying assumptions across all anticipated scenarios, inverse reinforcement learning provides inference on the short-term (local) rules governing long-term behavior policies by using properties of a Markov decision process. We use the computationally efficient linearly-solvable Markov decision process to learn the local rules governing collective movement for a simulation of the selfpropelled-particle (SPP) model and a data application for a captive guppy population. The estimation of the behavioral decision costs is done in a Bayesian framework with basis function smoothing. We recover the true costs in the SPP simulation and find the guppies value collective movement more than targeted movement toward shelter.


A FLEXIBLE SENSITIVITY ANALYSIS APPROACH FOR UNMEASURED CONFOUNDING WITH MULTIPLE TREATMENTS AND A BINARY OUTCOME WITH APPLICATION TO SEER-MEDICARE LUNG CANCER DATA

BY LIANGYUAN Hu1,a, JUNGAN ZOU2,b, CHENYANG GU3,c, JIAyi Ji4,d, MICHAEL LOPEZ5,e AND MINAL KALE6,f

1Department of Biostatistics and Epidemiology, Rutgers University, 4liangyuan.hu@rutgers.edu
2Department of Biostatistics, Columbia University, 5jzi3183@cumc.columbia.edu
3Department of Biostatistics, Brown University, 6chenyang.gu@alumni.brown.edu
4Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, 7jiayi.ji@mount Sinai.org
5Department of Mathematics, Skidmore College, 8mlopez1@skidmore.edu
6Department of Medicine, Icahn School of Medicine at Mount Sinai, 9minal.kale@mount Sinai.org

In the absence of a randomized experiment, a key assumption for drawing causal inference about treatment effects is the ignorable treatment assignment. Violations of the ignorability assumption may lead to biased treatment effect estimates. Sensitivity analysis helps gauge how causal conclusions will be altered in response to the potential magnitude of departure from the ignorability assumption. However, sensitivity analysis approaches for unmeasured confounding in the context of multiple treatments and binary outcomes are scarce. We propose a flexible Monte Carlo sensitivity analysis approach for causal inference in such settings. We first derive the general form of the bias introduced by unmeasured confounding, with emphasis on theoretical properties uniquely relevant to multiple treatments. We then propose methods to encode the impact of unmeasured confounding on potential outcomes and adjust the estimates of causal effects in which the presumed unmeasured confounding is removed. Our proposed methods embed nested multiple imputation within the Bayesian framework, which allow for seamless integration of the uncertainty about the values of the sensitivity parameters and the sampling variability as well as use of the Bayesian Additive Regression Trees for modeling flexibility. Expansive simulations validate our methods and gain insight into sensitivity analysis with multiple treatments. We use the SEER-Medicare data to demonstrate sensitivity analysis using three treatments for early stage nonsmall cell lung cancer. The methods developed in this work are readily available in the R package SAMTx.

REFERENCES


Big Data often presents as massive nonprobability samples. Not only is the selection mechanism often unknown but larger data volume amplifies the relative contribution of selection bias to total error. Existing bias adjustment approaches assume that the conditional mean structures have been correctly specified for the selection indicator or key substantive measures. In the presence of a reference probability sample, these methods rely on a pseudolikelihood method to account for the sampling weights of the reference sample, which is parametric in nature. Under a Bayesian framework, handling the sampling weights is an even bigger hurdle. To further protect against model misspecification, we expand the idea of double robustness such that more flexible nonparametric methods as well as Bayesian models can be used for prediction. In particular, we employ Bayesian additive regression trees which not only capture nonlinear associations automatically but permit direct quantification of the uncertainty of point estimates through its posterior predictive draws. We apply our method to sensor-based naturalistic driving data from the second Strategic Highway Research Program using the 2017 National Household Travel Survey as a benchmark.

REFERENCES


Key words and phrases. Big Data, nonprobability sample, quasi-randomization, prediction model, doubly robust, augmented inverse propensity weighting, Bayesian additive regression trees.


A SPARSE NEGATIVE BINOMIAL CLASSIFIER WITH COVARIATE ADJUSTMENT FOR RNA-SEQ DATA

BY TANBIN RAHMAN1,a, HSIN-EN HUANG2,d, YUJIA LI1,b, AN-SHUN TAI2,e, WEN-PING HSEIH2,f, COLLEEN A. McCLUNG3,g AND GEORGE TSENG1,c

1Department of Biostatistics, University of Pittsburgh, 2Department of Statistics, National Tsing Hua University, 3Department of Psychiatry, University of Pittsburgh.

Supervised machine learning methods have been increasingly used in biomedical research and clinical practice. In transcriptomic applications, RNA-seq data have become dominating and have gradually replaced traditional microarray, due to their reduced background noise and increased digital precision. Most existing machine learning methods are, however, designed for continuous intensities of microarray and are not suitable for RNA-seq count data. In this paper we develop a negative binomial model via generalized linear model framework with double regularization for gene and covariate sparsity to accommodate three key elements: adequate modeling of count data with overdispersion, gene selection and adjustment for covariate effect. The proposed sparse negative binomial classifier (snbClass) is evaluated in simulations and two real applications of multidisease postmortem brain tissue RNA-seq data and cervical tumor miRNA-seq data to demonstrate its superior performance in prediction accuracy and feature selection.

REFERENCES


Key words and phrases. Classification, count data, RNA-seq data, covariate adjustment.


KERNEL MACHINE AND DISTRIBUTED LAG MODELS FOR ASSESSING WINDOWS OF SUSCEPTIBILITY TO ENVIRONMENTAL MIXTURES IN CHILDREN’S HEALTH STUDIES

BY ANDER WILSON1,a, HSIAO-HSIEN LEON HSU2,b, YUEH-HSIU MATHILDA CHIU2,c, ROBERT O. WRIGHT2,d, ROSALIND J. WRIGHT2,e AND BRENT A. COULL3,f

1Department of Statistics, Colorado State University, aander.wilson@colostate.edu
2Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, bleon.hsu@mssm.edu, cmathilda.chiu@mssm.edu, drobert.wright@mssm.edu, erosalind.wright@mssm.edu
3Department of Biostatistics, Harvard T. H. Chan School of Public Health, fcoull@hsph.harvard.edu

Exposures to environmental chemicals during gestation can alter health status later in life. Most studies of maternal exposure to chemicals during pregnancy have focused on a single chemical exposure observed at high temporal resolution. Recent research has turned to focus on exposure to mixtures of multiple chemicals, generally observed at a single time point. We consider statistical methods for analyzing data on chemical mixtures that are observed at a high temporal resolution. As motivation, we analyze the association between exposure to four ambient air pollutants observed weekly throughout gestation and birth weight in a Boston-area prospective birth cohort. To explore patterns in the data, we first apply methods for analyzing data on: (1) a single chemical observed at high temporal resolution, and (2) a mixture measured at a single point in time. We highlight the shortcomings of these approaches for temporally-resolved data on exposure to chemical mixtures. Second, we propose a novel method, a Bayesian kernel machine regression distributed lag model (BKMR-DLM) that simultaneously accounts for nonlinear associations and interactions among time-varying measures of exposure to mixtures. BKMR-DLM uses a functional weight for each exposure that parameterizes the window of susceptibility corresponding to that exposure within a kernel machine framework that captures nonlinear and interaction effects of the multivariate exposure on the outcome. In a simulation study we show that the proposed method can better estimate the exposure-response function and, in high signal settings, can identify critical windows in time during which exposure has an increased association with the outcome. Applying the proposed method to the Boston birth cohort data, we find evidence of a negative association between organic carbon and birth weight and that nitrate modifies the organic carbon, elemental carbon, and sulfate exposure-response functions.

REFERENCES


Key words and phrases. Air pollution, chemical mixtures, children’s health, windows of susceptibility, distributed lag models, kernel machine regression.


DETECTING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES AND APPLICATION TO THE OREGON HEALTH INSURANCE EXPERIMENT

BY MICHAEL JOHNSON\textsuperscript{1,a}, JIONGYI CAO\textsuperscript{2,c} AND HYUNSEUNG KANG\textsuperscript{1,b}

\textsuperscript{1}Department of Statistics, University of Wisconsin-Madison, \textsuperscript{2}mwjohnson8@wisc.edu, \textsuperscript{b}hyunseung@stat.wisc.edu
\textsuperscript{2}Department of Statistics, University of Chicago, \textsuperscript{c}jiongyi@uchicago.edu

There is an increasing interest in estimating heterogeneity in causal effects in randomized and observational studies. However, little research has been conducted to understand effect heterogeneity in an instrumental variables study. In this work we present a method to estimate heterogeneous causal effects using an instrumental variable with matching. The method has two parts. The first part uses subject-matter knowledge and interpretable machine-learning techniques, such as classification and regression trees, to discover potential effect modifiers. The second part uses closed testing to test for statistical significance of each effect modifier while strongly controlling the familywise error rate. We apply this method on the Oregon Health Insurance Experiment, estimating the effect of Medicaid on the number of days an individual’s health does not impede their usual activities by using a randomized lottery as an instrument. Our method revealed Medicaid’s effect was most impactful among older, English-speaking, non-Asian males and younger, English-speaking individuals with, at most, a high school diploma or General Educational Development.

REFERENCES


Key words and phrases. Causal inference, complier average causal effect, heterogeneous treatment, instrumental variables, matching, machine learning, Oregon Health Insurance Experiment.


The arterial networks in the human brain, termed brain arterial networks or BANs, are complex arrangements of individual arteries, branching patterns, and interconnectivity. BANs play an essential role in characterizing and understanding brain physiology, and one would like tools for statistically analyzing the shapes of BANs. These tools include quantifying shape differences, comparing populations of subjects, and studying the effects of covariates on these shapes. This paper mathematically represents and statistically analyzes BAN shapes as elastic shape graphs. Each elastic shape graph consists of nodes, or points in 3D, connected by 3D curves, or edges, with arbitrary shapes. We develop a mathematical representation, a Riemannian metric and other geometrical tools, such as computations of geodesics, means, covariances, and PCA, for helping analyze BANs as elastic graphs. We apply this analysis to BANs after dividing them into four components—top, bottom, left, and right. The framework is then used to generate shape summaries of BANs from 92 subjects and study the effects of age and gender on shapes of BAN components. While gender effects require further investigation, we conclude that age has a clear, quantifiable effect on BAN shapes. Specifically, we find an increased variance in BAN shapes as age increases.


SPATIOTEMPORAL-TEXTUAL POINT PROCESSES FOR CRIME LINKAGE DETECTION

BY SHIXIANG ZHU\textsuperscript{a} AND YAO XIE\textsuperscript{b}

School of Industrial and Systems Engineering, Georgia Institute of Technology, \textsuperscript{a}shixiang.zhu@gatech.edu, \textsuperscript{b}Yao.xie@isye.gatech.edu

Crimes emerge out of complex interactions of human behaviors and situations. Linkages between crime incidents are highly complex. Detecting crime linkage, given a set of incidents, is a highly challenging task since we only have limited information, including text descriptions, incident times, and locations. In practice, there are very few labels. We propose a new statistical modeling framework for spatiotemporal-textual data and demonstrate its usage on crime linkage detection. We capture linkages of crime incidents via multivariate marked spatiotemporal Hawkes processes and treat embedding vectors of the free-text as marks of the incident, inspired by the notion of modus operandi (M.O.) in crime analysis. Numerical results, using real data, demonstrate the good performance of our method as well as reveals interesting patterns in the crime data: the joint modeling of space, time, and text information enhances crime linkage detection, compared with the state-of-the-art, and the learned spatial dependence from data can be useful for police operations.

REFERENCES


Key words and phrases. Hawkes processes with marks, crime linkage detection, variable selection.


MARKOV-MODULATED HAWKES PROCESSES FOR MODELING SPORADIC AND BURSTY EVENT OCCURRENCES IN SOCIAL INTERACTIONS

BY JING WU1,a OWEN G. WARD1,b, JAMES CURLEY2,d AND TIAN ZHENG1,c

1Department of Statistics, Columbia University, aju3233@columbia.edu, bowen.ward@columbia.edu, ctian.zheng@columbia.edu
2Department of Psychology, University of Texas at Austin, dcurley@utexas.edu

Modeling event dynamics is central to many disciplines. Patterns in observed social interaction events can be commonly modeled using point processes. Such social interaction event data often exhibit self-exciting, heterogeneous and sporadic trends which is challenging for conventional models. It is reasonable to assume that there exists a hidden state process that drives different event dynamics at different states. In this paper we propose a Markov modulated Hawkes process (MMHP) model for learning such a mixture of social interaction event dynamics and develop corresponding inference algorithms. Numerical experiments using synthetic data demonstrate that MMHP with the proposed estimation algorithms consistently recover the true hidden state process in simulations, while email data from a large university and data from an animal behavior study show that the procedure captures distinct event dynamics that reveal interesting social structures in the real data.

REFERENCES


Key words and phrases. Social interaction dynamics, self-exciting processes, heterogeneous point processes, latent Markov processes, Bayesian inference.


CONDITIONAL FUNCTIONAL CLUSTERING FOR LONGITUDINAL DATA WITH HETEROGENEOUS NONLINEAR PATTERNS

BY TIANHAO WANG\textsuperscript{a}, LEI YU\textsuperscript{b}, SUE E. LEURGANS\textsuperscript{c}, ROBERT S. WILSON\textsuperscript{d}, DAVID A. BENNETT\textsuperscript{e} AND PATRICIA A. BOYLE\textsuperscript{f}

Rush Alzheimer’s Disease Center, Rush University Medical Center, \textsuperscript{a}tianhao_wang@rush.edu, \textsuperscript{b}lei_yu@rush.edu, \textsuperscript{c}sue_e_leurgans@rush.edu, \textsuperscript{d}robert_s_wilson@rush.edu, \textsuperscript{e}david_a_bennett@rush.edu, \textsuperscript{f}patricia_boyle@rush.edu

In studies of cognitive aging, it is crucial to distinguish subtypes of longitudinal cognition change while accounting for the effects of given covariates. The longitudinal cognition trajectories and the covariate effects can both be nonlinear with heterogeneous shapes that do not follow a simple parametric form, where flexible functional methods are preferred. However, most functional clustering methods for longitudinal data do not allow controlling for the possible functional effects of covariates. Although traditional mixture-of-experts methods can include covariates and be extended to the functional setting, using nonlinear basis functions, satisfactory parsimonious functional methods required for robust functional coefficient estimation and clustering are still lacking. In this paper we propose a novel latent class functional mixed-effects model in which we assume the covariates have fixed functional effects, and the random curves follow a mixture of Gaussian processes that facilitates a model-based conditional clustering. A transformed penalized B-spline approach is employed for parsimonious modeling and robust model estimation. We propose a new iterative-REML method to choose the penalty parameters in heterogeneous data. The new method is applied to the latest data from the Religious Orders Study and Rush Memory and Aging Project, and four novel subtypes of cognitive changes are identified.

REFERENCES


Key words and phrases. Cohort study, functional mixed-effects model, latent classes, mixture Gaussian processes, penalized B-splines, ROSMAP studies.


IMPACT EVALUATION OF THE LAPD COMMUNITY SAFETY PARTNERSHIP

BY SYDNEY KAHMANN\textsuperscript{1,a}, ERIN HARTMAN\textsuperscript{2,b}, JORJA LEAP\textsuperscript{3,c} and P. JEFFREY BRANTINGHAM\textsuperscript{4,d}

\textsuperscript{1}Department of Statistics, University of California, Los Angeles, \textsuperscript{a}skahmann@ucla.edu  
\textsuperscript{2}Department of Political Science, University of California, Berkeley, \textsuperscript{b}ekhartman@berkeley.edu  
\textsuperscript{3}Department of Social Welfare, University of California, Los Angeles, \textsuperscript{c}jleap@ucla.edu  
\textsuperscript{4}Department of Anthropology, University of California, Los Angeles, \textsuperscript{d}branting@ucla.edu

In 2011, the Los Angeles Police Department (LAPD), in conjunction with other governmental and nonprofit groups, launched the Community Safety Partnership (CSP) in several public housing developments in Los Angeles. Following a relationship-based policing model, officers were assigned to work collaboratively with community members to reduce crime and build trust. However, evaluating the causal impact of this policy intervention is difficult, given the notable differences between communities where CSP was implemented and the surrounding communities in South Los Angeles. In this paper we use a novel data set, based on the LAPD’s reported crime incidents and calls-for-service, to evaluate the effectiveness of this program via augmented synthetic control models, a cutting-edge method for policy evaluation. We perform falsification analyses to evaluate the robustness of the results. In the public housing developments where it was first deployed, we find that CSP exhibited modest but statistically insignificant reductions in reported violent crime incidents, shots fired and violent crime calls-for-service, and Part I reported crime incidents. We do not find evidence of crime displacement from CSP regions to neighboring control regions.

REFERENCES


\textit{Key words and phrases.} Gang violence, community policing, time-series cross-sectional data, augmented synthetic control method, policy evaluation, causal inference.


LAPD NEWS RELEASE (2015). LAPD’s Community Safety Partnership program NR15021SF.


STREETER, K. (2014). In Jordan Downs housing project, police are forging a new relationship.

HIGHER CRITICISM FOR DISCRIMINATING WORD-FREQUENCY TABLES
AND AUTHORSHIP ATTRIBUTION

BY ALON KIPNIS

Department of Statistics, Stanford University, kipnisal@stanford.edu

We adapt the higher criticism (HC) goodness-of-fit test to measure the closeness between word-frequency tables. We apply this measure to authorship attribution challenges, where the goal is to identify the author of a document using other documents whose authorship is known. The method is simple yet performs well without handcrafting and tuning, reporting accuracy at the state-of-the-art level in various current challenges. As an inherent side effect, the HC calculation identifies a subset of discriminating words. In practice, the identified words have low variance across documents belonging to a corpus of homogeneous authorship. We conclude that in comparing the similarity of a new document and a corpus of a single author, HC is mostly affected by words characteristic of the author and is relatively unaffected by topic structure.

REFERENCES


Key words and phrases. Higher criticism, two-sample testing, nonparametric methods, unsupervised learning, feature selection, authorship attribution.


PEARSON, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **50** 157–175.


Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction

Bradley Efron

We live in a new age for statistical inference, where modern scientific technology such as microarrays and fMRI machines routinely produce thousands and sometimes millions of parallel data sets, each with its own estimation or testing problem. Doing thousands of problems at once is more than repeated application of classical methods. Taking an empirical Bayes approach, Bradley Efron, inventor of the bootstrap, shows how information accrues across problems in a way that combines Bayesian and frequentist ideas. Estimation, testing, and prediction blend in this framework, producing opportunities for new methodologies of increased power. New difficulties also arise, easily leading to flawed inferences. This book takes a careful look at both the promise and pitfalls of large-scale statistical inference, with particular attention to false discovery rates, the most successful of the new statistical techniques. Emphasis is on the inferential ideas underlying technical developments, illustrated using a large number of real examples.