Real-time mechanistic Bayesian forecasts of COVID-19 mortality
Graham C. Gibson, Nicholas G. Reich and Daniel Sheldon 1801

Calibration of SpatioTemporal forecasts from citizen science urban air pollution data with sparse recurrent
neural networks ................................. Matthew Bonas and Stefano Castruccio 1820

Tracking hematopoietic stem cell evolution in a Wiskott–Aldrich clinical trial
Danilo Pellin, Luca Biasco, Serena Scala, Clelia Di Serio and Ernst C. Wit 1841

Bayesian modeling of interaction between features in sparse multivariate count data with application to
microbiome study ................................. Shuangjie Zhang, Yuning Shen, Irene A. Chen and Juhee Lee 1861

Probabilistic learning of treatment trees in cancer ................................. Tsung-Hung Yao, Zhenke Wu,
Karthik Bharath, Jinju Li and Veerabhadran Baladandayuthapani 1884

Sequentially valid tests for forecast calibration
Sebastian Arnold, Alexander Henni and Johanna F. Ziegel 1909

Bayesian additive regression trees for genotype by environment interaction models ............................. Danilo A. Sarti,
Estevão B. Prado, Alan N. Inglis, Antonia A. L. dos Santos, Catherine B. Hurley, Rafael A. Moral and Andrew C. Parnell 1936

The scalable birth–death MCMC algorithm for mixed graphical model learning with application to genomic
data integration................................. Nanwei Wang, Helene Massam, Xin Gao and Laurent Brioliais 1958

A Bayesian hierarchical model framework to quantify uncertainty of tropical cyclone precipitation forecasts
Stephen Walsh, Marco A. R. Ferreira, David Higdon and Stephanie Zick 1984

Joint point and variance estimation under a hierarchical Bayesian model for survey count data
Terrance D. Savitsky, Julie Gershunskaya and Mark Crankshaw 2002

Data-adaptive discriminative feature localization with statistically guaranteed interpretation
Ben Dai, Xiaotong Shen, Lin Yee Chen, Chunlin Li and Wei Pan 2019

Dynamic prediction of residual life with longitudinal covariates using long short-term memory networks
Grace Rhodes, Marie Davidian and Wenbin Lu 2039

Postelection analysis of presidential election/poll data
Jiming Jiang, Yuanyuan Li and Peter X. K. Song 2059

Log-Gaussian Cox process modeling of large spatial lightning data using spectral and Laplace
approximations ................................. Megan L. Gelsinger, Maryclare Griffin, David Matteson and Joseph Guinness 2078

Graph-aware modeling of brain connectivity networks
Yura Kim, Daniel Kessler and Elizaveta Levina 2095

Bayesian model selection: Application to the adjustment of fundamental physical constants
Olha Bodnar and Viktor Eriksson 2118

Leveraging population outcomes to improve the generalization of experimental results: Application to the
JTPA study ................................. Melody Huang, Naoki Egami, Erin Hartman and Luke Miratrix 2139

Estimating causal effects of HIV prevention interventions with interference in network-based studies among
people who inject drugs ................................. Tingfang Lee, Ashley L. Buchanan, Natalia V. Katenka,
Laura Forastiere, M. Elizabeth Halloran, Samuel R. Friedman and Georgios Nikolopoulos 2165

Using persistent homology topological features to characterize medical images: Case studies on lung and
brain cancers ......................................... Chul Moon, Qimei Li and Guanghua Xiao 2192

Bayesian combinatorial MultiStudy factor analysis
Isabella N. Grabski, Roberta De Vito, Lorenzo Trippa and Giovanni Parmigiani 2212

continued
Using proxies to improve forecast evaluation ....................... HAJO HOLZMANN AND BERNHARD KLAR 2236
Bayesian nonparametric mixture modeling for temporal dynamics of gender stereotypes 
MARIA DE IORIO, STEFANO FAVARO, ALESSANDRA GUGLIELMI AND LIFENG YE 2256
The Bayesian nested lasso for mixed frequency regression models 
SATYAJIT GHOSH, KSITI KHARE AND GEORGE MICHAELIDIS 2279
Spatial quantile autoregression for season within year daily maximum temperature data 
JORGE CASTILLO-MATEO, JESÚS ASÍN, ANA C. CEBRIÁN, ALAN E. GELFAND AND JESÚS ABUARRA 2305
A dynamic screening algorithm for hierarchical binary marketing data 
YIMEI FAN, YUAN LIAO, ILYA O. RYZHOV AND KUNPENG ZHANG 2326
Penalized estimating equations for generalized linear models with multiple imputation 
YANG LI, HAOYU YANG, HAOCHEN YU, HANWEN HUANG AND YE SHEN 2345
Subbotin graphical models for extreme value dependencies with applications to functional neuronal connectivity 
ANDERSEN CHANG AND GENEVERA I. ALLEN 2364
Doubly-online changepoint detection for monitoring health status during sports activities 
MATTIA STIVAL, MAURO BERNARDI AND PETROS DELLAPORTAS 2387
Signal-noise ratio of genetic associations and statistical power of SNP-set tests 
HONG ZHANG, MING LIU, JIA SHUN JIN AND ZHEYANG WU 2410
Evaluating the use of generalized dynamic weighted ordinary least squares for individualized HIV treatment 
strategies ......................... LARRY DONG, ÉRICA E. M. MOODIE, LAURA VILLAIN AND RODOLPHE THIÉBAUT 2432
Imputation scores .......................... JEFFREY NAF, META-LINA SPOHN, LORIS MICHEL AND NICOLAI MEINSHAUSEN 2452
An efficient doubly-robust imputation framework for longitudinal dropout, with an application to an Alzheimer’s clinical trial ............................................................... YUJI QIU AND KAREN MEISSER 2473
A Bayesian growth mixture model for complex survey data: Clustering postdisaster PTSD trajectories 
REBECCA ANTHOPOLOS, QIXUAN CHEN, JOSEPH SEDRANSK, MARY THOMPSON, 
GANG MENG AND SANDRO GALEA 2494
Estimating HIV epidemics for subnational areas . . LE BAO, XIAOYUE NIU, MARY MAHY AND PETER D. GHYS 2515
Joint modeling of playing time and purchase propensity in massively multiplayer online role-playing games using crossed random effects ......................... TRAMBK BANERJEE, PENG LIU, GOURAB MUKHERJEE, 
SHANTANU DUTTA AND HAI CHE 2533
Structure learning for zero-inflated counts with an application to single-cell RNA sequencing data 
THI KIM HUE NGUYEN, KEOI VAN DEN BERGE, MONICA CHIOGNA AND DAVIDE RISSO 2555
Bayesian inference and dynamic prediction for multivariate longitudinal and survival data 
HAOTIAN ZOU, DONGLIN ZENG, LUO XIAO AND SHENG LUO 2574
Estimating GARCH(1, 1) in the presence of missing data 
DAMIEN C. H. WEE, FENG CHEN AND WILLIAM T. M. DUNSMUIR 2596
SNIP. An adaptation of sorted neighborhood methods for deduplicating pedigree data 
THEODORE HUANG, MATTHEW PLOENZKE AND DANIELLE BRAUN 2619
A horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in 
brain imaging ....................... FRANCESCO DENTI, RICCARDO AZEVEDO, CHELSIE LO, DAMIAN G. WHEELER, 
SUNIL GANDHI, MICHELE GUINDANI AND BABAK SHAHBABA 2639
Using predictability to improve matching of urban locations in Philadelphia 
COLMAN HUMPHREY, RYAN GROSS, DYLAN S. SMALL AND SHANE T. JENSEN 2659
A semiparametric promotion time cure model with support vector machine 
SUVRAN PAL AND WISDOM ASELISEWINE 2680
Corrigendum
Modeling biomarker ratios with gamma distributed components ......................... MATTHIAS SCHMID 2700
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: Michael Kosorok, Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, Chapel Hill, NC 27599, USA

President-Elect: Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA

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

Executive Secretary: Peter Hoff, Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA

Treasurer: Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, 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 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: Moulinath Banerjee, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

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

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 17, Number 3, September 2023. 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 © 2023 by the Institute of Mathematical Statistics
Printed in the United States of America
REAL-TIME MECHANISTIC BAYESIAN FORECASTS OF COVID-19 MORTALITY

BY GRAHAM C. GIBSON¹,a, NICHOLAS G. REICH¹,b AND DANIEL SHELDON²,c

¹ School of Public Health and Health Sciences, University of Massachusetts Amherst, a gcgibson@lanl.gov, b nick@umass.edu
² College of Information and Computer Sciences, University of Massachusetts Amherst, c sheldon@cs.umass.edu

The COVID-19 pandemic emerged in late December 2019. In the first six months of the global outbreak, the U.S. reported more cases and deaths than any other country in the world. Effective modeling of the course of the pandemic can help assist with public health resource planning, intervention efforts, and vaccine clinical trials. However, building applied forecasting models presents unique challenges during a pandemic. First, case data available to models in real time represent a nonstationary fraction of the true case incidence due to changes in available diagnostic tests and test-seeking behavior. Second, interventions varied across time and geography leading to large changes in transmissibility over the course of the pandemic. We propose a mechanistic Bayesian model that builds upon the classic compartmental susceptible–exposed–infected–recovered (SEIR) model to operationalize COVID-19 forecasting in real time. This framework includes nonparametric modeling of varying transmission rates, nonparametric modeling of case and death discrepancies due to testing and reporting issues, and a joint observation likelihood on new case counts and new deaths; it is implemented in a probabilistic programming language to automate the use of Bayesian reasoning for quantifying uncertainty in probabilistic forecasts. The model has been used to submit forecasts to the U.S. Centers for Disease Control through the COVID-19 Forecast Hub under the name MechBayes. We examine the performance relative to a baseline model as well as alternate models submitted to the forecast hub. Additionally, we include an ablation test of our extensions to the classic SEIR model. We demonstrate a significant gain in both point and probabilistic forecast scoring measures using MechBayes, when compared to a baseline model, and show that MechBayes ranks as one of the top two models out of nine which regularly submitted to the COVID-19 Forecast Hub for the duration of the pandemic, trailing only the COVID-19 Forecast Hub ensemble model of which which MechBayes is a part.

REFERENCES


Key words and phrases. COVID-19 mortality forecasting, compartmental models, Bayesian differential equations.

CALIBRATION OF SPATIOTEMPORAL FORECASTS FROM CITIZEN SCIENCE URBAN AIR POLLUTION DATA WITH SPARSE RECURRENT NEURAL NETWORKS

BY MATTHEW BONAS\textsuperscript{a} AND STEFANO CASTRUCCIO\textsuperscript{b}

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, \textsuperscript{a}mbonas@nd.edu, \textsuperscript{b}scastruc@nd.edu

With their continued increase in coverage and quality, data collected from personal air quality monitors has become an increasingly valuable tool to complement existing public health monitoring systems over urban areas. However, the potential of using such “citizen science data” for automatic early warning systems is hampered by the lack of models able to capture the high-resolution, nonlinear spatiotemporal features stemming from local emission sources such as traffic, residential heating and commercial activities. In this work we propose a machine-learning approach to forecast high-frequency spatial fields which has two distinctive advantages from standard neural network methods in time: (1) sparsity of the neural network via a spike-and-slab prior and (2) a small parametric space. The introduction of stochastic neural networks generates additional uncertainty, and in this work we propose a fast approach for ensure that the forecast is correctly assessed (calibration), both marginally and spatially. We focus on assessing exposure to urban air pollution in San Francisco, and our results suggest an improvement of over 58% in the mean squared error over standard time-series approach with a calibrated forecast for up to five days.

REFERENCES


Key words and phrases. Air pollution, spatiotemporal model, citizen science data, recurrent neural network.


Hematopoietic stem cells (HSC) are the cells that give rise to all other blood cells and, as such, they are crucial in the healthy development of individuals. Wiskott–Aldrich Syndrome (WAS) is a severe disorder affecting the regulation of hematopoietic cells and is caused by mutations in the WASP gene. We consider data from a revolutionary gene therapy clinical trial, where HSC harvested from three WAS patients’ bone marrow have been edited and corrected using viral vectors. Upon reinfusion into the patient, the HSC multiply and differentiate into other cell types. The aim is to unravel the cell multiplication and cell differentiation process, which has until now remained elusive. This paper models the replenishment of blood lineages resulting from corrected HSC via a multivariate, density-dependent Markov process and develops an inferential procedure to estimate the dynamic parameters, given a set of temporally sparsely observed trajectories. Starting from the master equation, we derive a system of nonlinear differential equations for the evolution of the first- and second-order moments over time. We use these moment equations in a generalized method-of-moments framework to perform inference. The performance of our proposal has been evaluated by considering different sampling scenarios and measurement errors of various strengths using a simulation study. We also compared it to another state-of-the-art approach and found that our method is statistically more efficient. By applying our method to the WAS gene therapy data, we found strong evidence for a myeloid-based developmental pathway of hematopoietic cells where fates of lymphoid and myeloid cells remain coupled, even after the loss of erythroid potential. All code used in this manuscript can be found in the online Supplementary Material, and the latest version of the code is available at https://github.com/dp3ll1n/SLCDP_v1.0.

REFERENCES


BY SHUANGJIE ZHANG1,a, YUNING SHEN2,c, IRENE A. CHEN2,d AND JUHEE LEE1,b

1Department of Statistics, University of California Santa Cruz, a szhan209@ucsc.edu, b juheelee@soe.ucsc.edu
2Department of Chemical and Biomolecular Engineering, University of California Los Angeles, cyshen@chem.ucsb.edu, direneachen@ucla.edu

Many statistical methods have been developed for the analysis of microbial community profiles, but due to the complexity of typical microbiome measurements, inference of interactions between microbial features remains challenging. We develop a Bayesian zero-inflated rounded log-normal kernel method to model interaction between microbial features in a community using multivariate count data in the presence of covariates and excess zeros. The model carefully constructs the interaction structure by imposing joint sparsity on the covariance matrix of the kernel and obtains a reliable estimate of the structure with a small sample size. The model also includes zero inflation to account for excess zeros observed in data and infers differential abundance of microbial features associated with covariates through log-linear regression. We provide simulation studies and real data analysis examples to demonstrate the developed model. Comparison of the model to a simpler model and popular alternatives in simulation studies shows that, in addition to an added and important insight on the feature interaction, it yields superior parameter estimates and model fit in various settings.

REFERENCES


Key words and phrases. Covariance matrix, differential abundance, factor model, joint sparsity, multivariate count data, rounded kernel model, zero inflation.


PROBABILISTIC LEARNING OF TREATMENT TREES IN CANCER

BY TSUNG-HUNG YAO$^{1,a}$, ZHENKE WU$^{1,b}$, KARTHIK BHARATH$^{2,e}$, JINJU LI$^{1,c}$ AND VEERABHADRAN BALADANDAYUTHAPANI$^{1,d}$

$^1$Department of Biostatistics, University of Michigan at Ann Arbor, $^a$yaots@umich.edu, $^b$zhenkewu@umich.edu, $^c$lijinju@umich.edu, $^d$veerab@umich.edu

$^2$School of Mathematical Sciences, University of Nottingham, $^e$Karthik.Bharath@nottingham.ac.uk

Accurate identification of synergistic treatment combinations and their underlying biological mechanisms is critical across many disease domains, especially cancer. In translational oncology research, preclinical systems, such as patient-derived xenografts (PDX), have emerged as a unique study design evaluating multiple treatments administered to samples from the same human tumor implanted into genetically identical mice. In this paper we propose a novel Bayesian probabilistic tree-based framework for PDX data to investigate the hierarchical relationships between treatments by inferring treatment cluster trees, referred to as treatment trees ($R_x$-tree). The framework motivates a new metric of mechanistic similarity between two or more treatments, accounting for inherent uncertainty in tree estimation; treatments with a high estimated similarity have potentially high mechanistic synergy. Building upon Dirichlet diffusion trees, we derive a closed-form marginal likelihood, encoding the tree structure, which facilitates computationally efficient posterior inference via a new two-stage algorithm. Simulation studies demonstrate superior performance of the proposed method in recovering the tree structure and treatment similarities. Our analyses of a recently collated PDX dataset produce treatment similarity estimates that show a high degree of concordance with known biological mechanisms across treatments in five different cancers. More importantly, we uncover new and potentially effective combination therapies that confer synergistic regulation of specific downstream biological pathways for future clinical investigations. Our accompanying code, data, and shiny application for visualization of results are available at: https://github.com/bayesrx/RxTree.

REFERENCES


Key words and phrases. Approximate Bayesian computation, Dirichlet diffusion trees, patient derived xenograft, precision medicine, tree-based clustering.
**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.

**Migraine headache:** A common, severe, pulsating headache that can be associated with nausea, sensitivity to light, and sensitivity to sound.
Yuan, Y., Wen, W., Yost, S. E., Xing, Q., Yan, J., Han, E. S., Mortimer, J. and Yim, J. H. (2019). Combination therapy with BYL719 and LEE011 is synergistic and causes a greater suppression of p-S6 in triple negative breast cancer. Sci. Rep. 9 7509.


SEQUENTIALLY VALID TESTS FOR FORECAST CALIBRATION

BY SEBASTIAN ARNOLD\textsuperscript{a}, ALEXANDER HENZI\textsuperscript{b} AND JOHANNA F. ZIEGEL\textsuperscript{c}

Institute of Mathematical Statistics and Actuarial Science, University of Bern, \textsuperscript{a}sebastian.arnold@stat.unibe.ch, \textsuperscript{b}alexander.henzi@stat.unibe.ch, \textsuperscript{c}johanna.ziegel@stat.unibe.ch

Forecasting and forecast evaluation are inherently sequential tasks. Predictions are often issued on a regular basis, such as every hour, day, or month, and their quality is monitored continuously. However, the classical statistical tools for forecast evaluation are static, in the sense that statistical tests for forecast calibration are only valid if the evaluation period is fixed in advance. Recently, e-values have been introduced as a new, dynamic method for assessing statistical significance. An e-value is a nonnegative random variable with expected value, at most, one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a conservative p-value. Since they naturally lead to statistical tests which are valid under optional stopping, e-values are particularly suitable for sequential forecast evaluation. This article proposes e-values for testing probabilistic calibration of forecasts which is one of the most important notions of calibration. The proposed methods are also more generally applicable for sequential goodness-of-fit testing. We demonstrate in a simulation study that the e-values are competitive in terms of power, when compared to extant methods which do not allow for sequential testing. In this context we introduce test power heat matrices, a graphical tool to compactly visualize results of simulation studies on test power. In a case study we show that the e-values provide important and new useful insights in the evaluation of probabilistic weather forecasts.

REFERENCES


Key words and phrases. Forecast calibration, probabilistic calibration, rank histogram, e-value, sequential inference.


BAYESIAN ADDITIVE REGRESSION TREES FOR GENOTYPE BY ENVIRONMENT INTERACTION MODELS

BY DANILO A. SARTI\textsuperscript{1,a}, ESTEVÃO B. PRADO\textsuperscript{1,2,b}, ALAN N. INGLIS\textsuperscript{1,2,c}, ANTÔNIA A. L. DOS SANTOS\textsuperscript{1,d}, CATHERINE B. HURLEY\textsuperscript{1,e}, RAFAEL A. MORAL\textsuperscript{1,f} AND ANDREW C. PARNELL\textsuperscript{1,2,g}

\textsuperscript{1} Insight Centre for Data Analytics, Maynooth University
\textsuperscript{2} Hamilton Institute, Department of Mathematics and Statistics, Maynooth University, \texttt{daniloasarti@gmail.com}, \texttt{estevao.prado@hotmail.com}, \texttt{alan.inglis@mu.ie}, \texttt{antonia.lemosdossantos.2020@mumail.ie}, \texttt{catherine.hurley@mu.ie}, \texttt{rafael.deandrademoral@mu.ie}, \texttt{Andrew.parnell@mu.ie}

We propose a new class of models for the estimation of genotype by environment (GxE) interactions in plant-based genetics. Our approach, named AMBARTI, uses semiparametric Bayesian additive regression trees to accurately capture marginal genotypic and environment effects along with their interaction in a cut Bayesian framework. We demonstrate that our approach is competitive or superior to similar models widely used in the literature via both simulation and a real world dataset. Furthermore, we introduce new types of visualisation to properly assess both the marginal and interactive predictions from the model. An R package that implements our approach is also available at \url{https://github.com/ebprado/ambarti}.

REFERENCES


\textit{Key words and phrases}. Bayesian nonparametric regression, Bayesian additive regression trees, additive main effects multiplicative interactions model, genotype-by-environment interactions.


Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications that allow the study of biological mechanisms at an unprecedented depth and scale. A large amount of genomic data is now distributed through consortia like The Cancer Genome Atlas (TCGA), where specific types of biological information on specific type of tissue or cell are available. In cancer research the challenge is now to perform integrative analyses of high-dimensional multiomic data with the goal to better understand genomic processes that correlate with cancer outcomes, for example, elucidate gene networks that discriminate a specific cancer subgroups (cancer subtyping) or discovering gene networks that overlap across different cancer types (pan-cancer studies). In this paper we propose a novel mixed graphical model approach to analyze multiomic data of different types (continuous, discrete and count) and perform model selection by extending the birth–death MCMC (BDMCMC) algorithm initially proposed by Stephens (Ann. Statist. 28 (2000) 40–74) and later developed by Mohammadi and Wit (Bayesian Anal. 10 (2015) 109–138). Using simulations, we compare the performance of our method to the LASSO method and the standard BDMCMC method and find that our method is superior in terms of both computational efficiency and the accuracy of the model selection results. Finally, an application to the TCGA breast cancer data shows that integrating genomic information at different levels (mutation and expression data) leads to better subtyping of breast cancers.
Tropical cyclones present a serious threat to many coastal communities around the world. Many numerical weather prediction models provide deterministic forecasts with limited measures of their forecast uncertainty. Standard postprocessing techniques may struggle with extreme events or use a 30-day training window that will not adequately characterize the uncertainty of a tropical cyclone forecast. We propose a novel approach that leverages information from past storm events, using a hierarchical model to quantify uncertainty in the spatial correlation parameters of the forecast errors (modeled as Gaussian processes) for a numerical weather prediction model. This approach addresses a massive data problem by implementing a drastic dimension reduction through the assumption that the MLE and Hessian matrix represent all useful information from each tropical cyclone. From this, simulated forecast errors provide uncertainty quantification for future tropical cyclone forecasts. We apply this method to the North American Mesoscale model forecasts and use observations based on the Stage IV data product for 47 tropical cyclones between 2004 and 2017. For an incoming storm, our hierarchical framework combines the forecast from the North American Mesoscale model with the information from previous storms to create 95% and 99% prediction maps of rain. For six test storms from 2018 and 2019, these maps provide appropriate probabilistic coverage of observations. We show evidence from the log scoring rule that the proposed hierarchical framework performs best among competing methods.

REFERENCES


JOINT POINT AND VARIANCE ESTIMATION UNDER A HIERARCHICAL BAYESIAN MODEL FOR SURVEY COUNT DATA

BY TERRANCE D. SAVITSKY¹,a, JULIE GERSHUNSKAYA²,b AND MARK CRANKSHAW²,c

¹Office of Survey Methods Research, U.S. Bureau of Labor Statistics. aSavitsky.Terrance@bls.gov
²OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics. bGershunskaya.Julie@bls.gov, cCrankshaw.Mark@bls.gov

We propose a novel Bayesian framework for the joint modeling of survey point and variance estimates for count data. The approach incorporates an induced prior distribution on the modeled true variance that sets it equal to the generating variance of the point estimate, a key property more readily achieved for continuous data response type models. Our count data model formulation allows the input of domains at multiple resolutions (e.g., states, regions, nation) and simultaneously benchmarks modeled estimates at higher resolutions (e.g., states) to those at lower resolutions (e.g., regions) in a fashion that borrows more strength to sharpen our domain estimates at higher resolutions. We conduct a simulation study that generates a population of units within domains to produce ground truth statistics to compare to direct and modeled estimates performed on samples taken from the population where we show improved reductions in error across domains. The model is applied to the job openings variable and other data items published in the Job Openings and Labor Turnover Survey administered by the U.S. Bureau of Labor Statistics.

REFERENCES


Key words and phrases. Bayesian hierarchical models, small area estimation, count data, Stan.

DATA-ADAPTIVE DISCRIMINATIVE FEATURE LOCALIZATION WITH STATISTIALLY GUARANTEED INTERPRETATION

BY BEN DAI1,a, XIATONG SHEN2,b, LIN YEE CHEN3,d, CHUNLIN LI2,c AND WEI PAN4,e

1Department of Statistics, The Chinese University of Hong Kong, a bendai@cuhk.edu.hk
2School of Statistics, University of Minnesota, b xshen@umn.edu, c li000007@umn.edu
3Lillehei Heart Institute and Cardiovascular Division, University of Minnesota, d chenx484@umn.edu
4Division of Biostatistics, University of Minnesota, e panxx014@umn.edu

In explainable artificial intelligence, discriminative feature localization is critical to reveal a black-box model’s decision-making process from raw data to prediction. In this article we use two real datasets, the MNIST handwritten digits and MIT-BIH electrocardiogram (ECG) signals, to motivate key characteristics of discriminative features, namely, adaptiveness, predictive importance and effectiveness. Then we develop a localization framework, based on adversarial attacks, to effectively localize discriminative features. In contrast to existing heuristic methods, we also provide a statistically guaranteed interpretability of the localized features by measuring a generalized partial $R^2$. We apply the proposed method to the MNIST dataset and the MIT-BIH dataset with a convolutional autoencoder. In the first, the compact image regions localized by the proposed method are visually appealing. Similarly, in the second, the identified ECG features are biologically plausible and consistent with cardiac electrophysiological principles while locating subtle anomalies in a QRS complex that may not be discernible by the naked eye. Overall, the proposed method compares favorably with state-of-the-art competitors. Accompanying this paper is a Python library dnn-locate that implements the proposed approach.

REFERENCES


Key words and phrases. Explainable artificial intelligence, discriminative features, localization, generalized partial $R^2$, interpretability, regularization, deep learning.


DYNAMIC PREDICTION OF RESIDUAL LIFE WITH LONGITUDINAL COVARIATES USING LONG SHORT-TERM MEMORY NETWORKS

BY GRACE RHODES\textsuperscript{a}, MARIE DAVIDIAN\textsuperscript{b} AND WENBIN LU\textsuperscript{c}

Department of Statistics, North Carolina State University, \textsuperscript{a}gmrhodes@ncsu.edu, \textsuperscript{b}davidian@ncsu.edu, \textsuperscript{c}wlu4@ncsu.edu

Sepsis, a complex medical condition that involves severe infections with life-threatening organ dysfunction, is a leading cause of death worldwide. Treatment of sepsis is highly challenging. When making treatment decisions, clinicians and patients desire accurate predictions of mean residual life (MRL) that leverage all available patient information, including longitudinal biomarker data. Biomarkers are biological, clinical, and other variables reflecting disease progression that are often measured repeatedly on patients in the clinical setting. Dynamic prediction methods leverage accruing biomarker measurements to improve performance, providing updated predictions as new measurements become available. We introduce two methods for dynamic prediction of MRL using longitudinal biomarkers. In both methods, we begin by using long short-term memory networks (LSTMs) to construct encoded representations of the biomarker trajectories, referred to as “context vectors.” In our first method, the LSTM-GLM, we dynamically predict MRL via a transformed MRL model that includes the context vectors as covariates. In our second method, the LSTM-NN, we dynamically predict MRL from the context vectors using a feed-forward neural network. We demonstrate the improved performance of both proposed methods relative to competing methods in simulation studies. We apply the proposed methods to dynamically predict the restricted mean residual life (RMRL) of septic patients in the intensive care unit using electronic medical record data. We demonstrate that the LSTM-GLM and the LSTM-NN are useful tools for producing individualized, real-time predictions of RMRL that can help inform the treatment decisions of septic patients.

REFERENCES


Key words and phrases. Biomarker, dynamic prediction, electronic medical record, long short-term memory network, longitudinal data, MIMIC-III, neural network, residual life, sepsis, transformed mean residual life model.
POSTELECTION ANALYSIS OF PRESIDENTIAL ELECTION/POLL DATA

BY JIMING JIANG1,a, YUANYUAN LI1,b AND PETER X. K. SONG2,c

1Department of Statistics, University of California, Davis, 3jimjiang@ucdavis.edu, byyynli@ucdavis.edu
2Department of Biostatistics, University of Michigan, cpxsong@umich.edu

This paper concerns analyses of the 2016 and 2020 U.S. presidential election data, including the data of pre-election polls and the actual elections. Our analyses unveil statistical evidence of discrepancy between the polls and real elections that is consistent across these two elections. Specifically, the polls had consistently overestimated advantages of the Democratic candidates or, equivalently, underestimated the true population support of the Republican candidate, Donald Trump, in both elections. The analyses are stratified by state, reflecting the U.S. electoral college system by the means of small area estimation. We have found recurrent patterns suggesting that the polls have been underestimating the Republican candidate, especially in swing states of critical importance. Our findings also suggest an improvement of the 2020 polling methods to mitigate the size of underestimation. We show that a small-area model built upon the actual election data from one election can provide a better prediction than the poll-based projection to another election involving the same Republican candidate. Ranking of pollsters, based on prediction bias, using mixed model prediction is also considered.

REFERENCES


Key words and phrases. Empirical BLUP, measure of uncertainty, mixed-effects model, opinion polls, projection, small area estimation.


Lightning is a destructive and highly visible product of severe storms, yet there is still much to be learned about the conditions under which lightning is most likely to occur. The GOES-16 and GOES-17 satellites, launched in 2016 and 2018 by NOAA and NASA, collect a wealth of data regarding individual lightning strike occurrence and potentially related atmospheric variables. The acute nature and inherent spatial correlation in lightning data renders standard regression analyses inappropriate. Further, computational considerations are foregrounded by the desire to analyze the immense and rapidly increasing volume of lightning data. We present a new computationally feasible method that combines spectral and Laplace approximations in an EM algorithm, denoted SLEM, to fit the widely popular log-Gaussian Cox process model to large spatial point pattern datasets. In simulations we find SLEM is competitive with contemporary techniques in terms of speed and accuracy. When applied to two lightning datasets, SLEM provides better out-of-sample prediction scores and quicker runtimes, suggesting its particular usefulness for analyzing lightning data which tend to have sparse signals.

REFERENCES


Key words and phrases. Spatial point pattern, Laplace approximation, expectation-maximization.


NOAA, N. G. D. C. Data Announcement 88-MGG-02, Digital relief of the Surface of the Earth.


GRAPH-AWARE MODELING OF BRAIN CONNECTIVITY NETWORKS

BY YURA KIM\textsuperscript{a}, DANIEL KESSLER\textsuperscript{b} AND ELIZAVETA LEVINA\textsuperscript{c}

Department of Statistics, University of Michigan, \textsuperscript{a}kimyr@umich.edu, \textsuperscript{b}kesslerd@umich.edu, \textsuperscript{c}elevina@umich.edu

Functional connections in the brain are frequently represented by weighted networks, with nodes representing locations in the brain and edges representing the strength of connectivity between these locations. One challenge in analyzing such data is that inference at the individual edge level is not particularly biologically meaningful; interpretation is more useful at the level of so-called functional systems or groups of nodes and connections between them; this is often called “graph-aware” inference in the neuroimaging literature. However, pooling over functional regions leads to significant loss of information and lower accuracy. Another challenge is correlation among edge weights within a subject which makes inference based on independence assumptions unreliable. We address both of these challenges with a linear mixed effects model, which accounts for functional systems and for edge dependence, while still modeling individual edge weights to avoid loss of information. The model allows for comparing two populations, such as patients and healthy controls, both at the functional regions level and at individual edge level, leading to biologically meaningful interpretations. We fit this model to resting state fMRI data on schizophrenic patients and healthy controls, obtaining interpretable results consistent with the schizophrenia literature.

REFERENCES


Key words and phrases. Neuroimaging, functional MRI, network analysis.


BAYESIAN MODEL SELECTION: APPLICATION TO THE ADJUSTMENT OF FUNDAMENTAL PHYSICAL CONSTANTS

BY OLHA BODNAR¹,a AND VIKTOR ERIKSSON²,b

¹Unit of Statistics, School of Business, Örebro University, a olha.bodnar@oru.se
²Department of Statistics, Uppsala University, b viktor.eriksson@statistik.uu.se

A method originally suggested by Raymond Birge, using what came to be known as the Birge ratio, has been widely used in metrology and physics for the adjustment of fundamental physical constants, particularly in the periodic reevaluation carried out by the Task Group on Fundamental Physical Constants of CODATA (the Committee on Data of the International Science Council). The method involves increasing the reported uncertainties by a multiplicative factor large enough to make the measurement results mutually consistent. An alternative approach, predominant in the meta-analysis of medical studies, involves inflating the reported uncertainties by combining them, using the root sum of squares, with a sufficiently large constant (often dubbed dark uncertainty) that is estimated from the data.

In this contribution we establish a connection between the method based on the Birge ratio and the location-scale model, which allows one to combine the results of various studies, while the additive adjustment is reviewed in the usual context of random-effects models. Framing these alternative approaches as statistical models facilitates a quantitative comparison of them using statistical tools for model comparison. The intrinsic Bayes factor (IBF) is derived for the Berger and Bernardo reference prior, and then it is used to select a model for a set of measurements of the Newtonian constant of gravitation ("Big G") to estimate a consensus value for this constant and to evaluate the associated uncertainty. Our empirical findings support the method based on the Birge ratio. The same conclusion is reached when the IBF corresponding to the Jeffreys prior is used and also when the comparison is based on the Akaike information criterion (AIC). Finally, the results of a simulation study indicate that the suggested procedure for model selection provides clear guidance, even when the data comprise only a small number of measurements.

REFERENCES


Key words and phrases. Intrinsic Bayes factor, Birge ratio method, location-scale model, random-effects model, reference prior, meta-analysis, interlaboratory comparison study, Newtonian constant of gravitation.
LEVERAGING POPULATION OUTCOMES TO IMPROVE THE GENERALIZATION OF EXPERIMENTAL RESULTS: APPLICATION TO THE JTPA STUDY

BY MELODY HUANG1,a, NAOKI EGAMI2,b, ERIN HARTMAN3,c AND LUKE MIRATRIX4,d

1Department of Statistics, University of California, Berkeley, 2melodyhuang@berkeley.edu
2Department of Political Science, Columbia University, 3naoki.egami@columbia.edu
3Departments of Political Science and Statistics, University of California, Berkeley, 4ekhartman@berkeley.edu
4Graduate School of Education, Harvard University, 5lmiratrix@g.harvard.edu

Generalizing causal estimates in randomized experiments to a broader target population is essential for guiding decisions by policymakers and practitioners in the social and biomedical sciences. While recent papers have developed various weighting estimators for the population average treatment effect (PATE), many of these methods result in large variance because the experimental sample often differs substantially from the target population and estimated sampling weights are extreme. We investigate this practical problem motivated by an evaluation study of the Job Training Partnership Act (JTPA), where we examine how well we can generalize the causal effect of job training programs beyond a specific population of economically disadvantaged adults and youths. In particular, we propose post-residualized weighting in which we use the outcome measured in the observational population data to build a flexible predictive model (e.g., with machine learning) and residualize the outcome in the experimental data before using conventional weighting methods. We show that the proposed PATE estimator is consistent under the same assumptions required for existing weighting methods, importantly without assuming the correct specification of the predictive model. We demonstrate the efficiency gains from this approach through our JTPA application: we find a reduction of between 5% and 25% in variance.

REFERENCES


Key words and phrases. Generalizability, external validity, randomized experiments, causal inference.


ESTIMATING CAUSAL EFFECTS OF HIV PREVENTION INTERVENTIONS WITH INTERFERENCE IN NETWORK-BASED STUDIES AMONG PEOPLE WHO INJECT DRUGS

BY TINGFANG LEE¹,a, ASHLEY L. BUCHANAN¹,b, NATALLIA V. KATENKA²,c, LAURA FORASTIERE³,d, M. ELIZABETH HALLORAN⁴,5,e, SAMUEL R. FRIEDMAN⁶,f AND GEORGIOS NIKOLOPOULOS⁷,g

¹Department of Pharmacy Practice, University of Rhode Island, a tingfanglee@uri.edu, b buchanan@uri.edu
²Department of Computer Science and Statistics, University of Rhode Island, c nkatenka@uri.edu
³Department of Biostatistics, Yale School of Public Health, d laura.forastiere@yale.edu
⁴Biostatistics, Bioinformatics, and Epidemiology Program, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center
⁵Department of Biostatistics, University of Washington, ebetz@fredhutch.org
⁶Department of Population Health, NYU Grossman School of Medicine, f Samuel.Friedman@nyulangone.org
⁷Medical School, University of Cyprus, g nikolopoulos.georgios@ucy.ac.cy

Evaluating causal effects in the presence of interference is challenging in network-based studies of hard-to-reach populations. Like many such populations, people who inject drugs (PWID) are embedded in social networks and one person’s treatment can affect the outcomes of others in the network. In our setting, the study design is observational with a nonrandomized network-based HIV prevention intervention. Information is available on each participant and their connections that confer possible HIV risk through injection and sexual behaviors. We considered two inverse probability weighted (IPW) estimators to quantify the population-level spillover effects of nonrandomized interventions on subsequent health outcomes. We demonstrated that these two IPW estimators are consistent, asymptotically normal, and derived a closed-form estimator for the asymptotic variance, while allowing for overlapping interference sets (groups of individuals in which the interference is assumed possible). A simulation study was conducted to evaluate the finite-sample performance of the estimators. We analyzed data from the Transmission Reduction Intervention Project which ascertained a network of PWID and their contacts in Athens, Greece, from 2013 to 2015. We evaluated the effects of community alerts on subsequent HIV risk behavior in this observed network, where the connections or links between participants were defined by using substances or having unprotected sex together. In the study, community alerts were distributed to inform people of recent HIV infections among individuals in close proximity in the observed network. The estimates of the risk differences for spillover, using either IPW estimator demonstrated a protective effect. The results suggest that HIV risk behavior could be mitigated by exposure to a community alert when an increased risk of HIV is detected in the network.

REFERENCES


Key words and phrases. Causal interference, interference/dissemination, network studies, people who use drugs, HIV/AIDS, inverse probability weights.


USING PERSISTENT HOMOLOGY TOPOLOGICAL FEATURES TO CHARACTERIZE MEDICAL IMAGES: CASE STUDIES ON LUNG AND BRAIN CANCERS

BY CHUL MOON¹,a, QIWEI LI²,b AND GUANGHUA XIAO³,c

¹Department of Statistical Science, Southern Methodist University, achulm@smu.edu
²Department of Mathematical Sciences, University of Texas at Dallas, bqiwei.li@utdallas.edu
³Quantitative Biomedical Research Center, Department of Population & Data Sciences and Department of Bioinformatics, University of Texas Southwestern Medical Center, cguanghua.xiao@utsouthwestern.edu

Tumor shape is a key factor that affects tumor growth and metastasis. This paper proposes a topological feature computed by persistent homology to characterize tumor progression from digital pathology and radiology images and examines its effect on the time-to-event data. The proposed topological features are invariant to scale-preserving transformation and can summarize various tumor shape patterns. The topological features are represented in functional space and used as functional predictors in a functional Cox proportional hazards model. The proposed model enables interpretable inference about the association between topological shape features and survival risks. Two case studies are conducted using consecutive 133 lung cancer and 77 brain tumor patients. The results of both studies show that the topological features predict survival prognosis after adjusting clinical variables, and the predicted high-risk groups have worse survival outcomes than the low-risk groups. Also, the topological shape features found to be positively associated with survival hazards are irregular and heterogeneous shape patterns which are known to be related to tumor progression.

REFERENCES


Key words and phrases. Topological data analysis, tumor shape, functional data analysis, survival analysis, Cox proportional hazards model.


American Journal of Translational Research 11 4491.

Revue Sci. 84 159–162. MR0017892

Med. Image Anal. 33 170–175.


https://doi.org/10.1214/12-AOAS1714SUPPA, https://doi.org/10.1214/12-AOAS1714SUPPPB

PloS ONE 10 e0143811.

Cancer 52 550–554. https://doi.org/10.1002/1097-0142(19830801)52:3<550::aid-cncr2820520327>3.0.co;2-c


European Radiology Experimental 2 1–8.


The Cancer Imaging Archive 11 1.


BAYESIAN COMBINATORIAL MULTISTUDY FACTOR ANALYSIS

BY ISABELLA N. GRABSKI¹,a, ROBERTA DE VITO²,b, LORENZO TRIPPA³,c AND GIOVANNI PARMI GIANI³,d

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, aisabellagrabski@g.harvard.edu
²Department of Biostatistics and Data Science Initiative, Brown University, broberta_devito@brown.edu
³Department of Data Science, Dana Farber Cancer Institute, cltrippa@jimmy.harvard.edu, dg@jimmy.harvard.edu

Mutations in the BRCA1 and BRCA2 genes are known to be highly associated with breast cancer. Identifying both shared and unique transcript expression patterns in blood samples from these groups can shed insight into if and how the disease mechanisms differ among individuals by mutation status, but this is challenging in the high-dimensional setting. A recent method, Bayesian multistudy factor analysis (BMSFA), identifies latent factors common to all studies (or equivalently, groups) and latent factors specific to individual studies. However, BMSFA does not allow for factors shared by more than one but less than all studies. This is critical in our context, as we may expect some but not all signals to be shared by BRCA1- and BRCA2-mutation carriers but not necessarily other high-risk groups. We extend BMSFA by introducing a new method, Tetris, for Bayesian combinatorial multistudy factor analysis which identifies latent factors that any combination of studies or groups can share. We model the subsets of studies that share latent factors with an Indian buffet process and offer a way to summarize uncertainty in the sharing patterns using credible balls. We test our method with an extensive range of simulations and showcase its utility not only in dimension reduction but also in covariance estimation. When applied to transcript expression data from high-risk families grouped by mutation status, Tetris reveals the features and pathways characterizing each group and the sharing patterns among them. Finally, we further extend Tetris to discover groupings of samples when group labels are not provided which can elucidate additional structure in these data.

REFERENCES


Key words and phrases. Multistudy analysis, unsupervised learning, Gibbs sampling, factor analysis, dimension reduction.


USING PROXIES TO IMPROVE FORECAST EVALUATION

BY HAJO HOLZMANN\textsuperscript{1,a} AND BERNHARD KLAR\textsuperscript{2,b}

\textsuperscript{1}Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, \textsuperscript{a}holzmann@mathematik.uni-marburg.de
\textsuperscript{2}Institut für Stochastik, Karlsruher Institut für Technologie (KIT), \textsuperscript{b}bernhard.klar@kit.edu

Comparative evaluation of forecasts of statistical functionals relies on comparing averaged losses of competing forecasts after the realization of the quantity $Y$, on which the functional is based, has been observed. Motivated by high-frequency finance, in this paper we investigate how proxies $\tilde{Y}$ for $Y$—say volatility proxies—which are observed together with $Y$ can be utilized to improve forecast comparisons. We extend previous results on robustness of loss functions for the mean to general moments and ratios of moments, and show in terms of the variance of differences of losses that using proxies will increase the power in comparative forecast tests. These results apply both to testing conditional as well as unconditional dominance. Finally, we numerically illustrate the theoretical results, both for simulated high-frequency data as well as for high-frequency log returns of several cryptocurrencies.

REFERENCES


\textit{Key words and phrases.} Cryptocurrencies, DM test, forecast comparison, volatility proxies.
BAYESIAN NONPARAMETRIC MIXTURE MODELING FOR TEMPORAL DYNAMICS OF GENDER STEREOTYPES

BY MARIA DE IORIO\textsuperscript{1,a}, STEFANO FAVARO\textsuperscript{2,b}, ALESSANDRA GUGLIELMI\textsuperscript{3,c} AND LIFENG YE\textsuperscript{4,d}

\textsuperscript{1} Yong Loo Lin School of Medicine, National University of Singapore; \textsuperscript{a}mdi@nus.edu.sg
\textsuperscript{2} Department of Economics and Statistics, Università di Torino and Collegio Carlo Alberto, \textsuperscript{b} stefano.favaro@unito.it
\textsuperscript{3} Department of Mathematics, Politecnico di Milano, \textsuperscript{c} alessandra.guglielmi@polimi.it
\textsuperscript{4} Department of Statistical Science, University College London, \textsuperscript{d} lifeng.ye.13@ucl.ac.uk

The study of temporal dynamics of gender and ethnic stereotypes is an important topic in many disciplines at the intersection between statistics and social sciences. In this paper we make use of word “embeddings,” a common tool in natural language processing and of Bayesian nonparametric mixture modeling for the analysis of temporal dynamics of gender stereotypes in adjectives and occupation over the 20th and 21st centuries in the United States. Our Bayesian nonparametric approach relies on a novel dependent Dirichlet process prior, and it allows for both dynamic density estimation and dynamic clustering of adjective embedding and occupation embedding biases in a hierarchical setting. Posterior inference is performed through a particle Markov chain Monte Carlo algorithm, which is simple and computationally efficient. An application to time-dependent data for adjective embedding bias and for occupation embedding bias shows that our approach enables the quantification of historical trends of gender stereotypes and hence allows to identify how specific adjectives and occupations have become more closely associated with a female rather than male over time.

REFERENCES


Key words and phrases. Autoregressive models, Bayesian nonparametrics, dependent Dirichlet processes, dynamic density estimation and clustering, gender stereotypes, mixture modeling, particle Markov chain Monte Carlo, word embeddings.


THE BAYESIAN NESTED LASSO FOR MIXED FREQUENCY REGRESSION MODELS

BY SATYAJIT GHOSH¹,a, KSITIJ KHARE²,b AND GEORGE MICHAILIDIS²,3,c

¹Division of Biometrics II, CDER, U.S. FDA, a satyajitghosh90@gmail.com
²Department of Statistics, University of Florida, b kdkhare@stat.ufl.edu
³Informatics Institute, University of Florida, c gmichail@ufl.edu

Even though many time series are sampled at different frequencies, their joint evolution is usually modeled and analyzed at a common low frequency. The mixed data sampling (MIDAS) framework was developed to enable joint modeling of mixed frequency temporally evolving data with GDP forecasting as a key motivating application. In this paper we develop a fully Bayesian method to jointly estimate both the appropriate lag as well as the regression coefficients in linear models wherein the response is measured at a lower frequency than the predictors. This is accomplished through a novel prior distribution, coined the Bayesian nested lasso (BNL), that leads to principled selection of the lag of the predictors, reduces the effective number of model parameters through sparsity induced by the lasso component and finally incorporates desirable decay patterns over time lags in the magnitude of the corresponding regression coefficients. Further, it is easy to obtain samples from the posterior distribution due to the closed form expressions for the conditional distributions of the model parameters. Numerical results, obtained from synthetic and macroeconomic data, illustrate the good performance of the proposed Bayesian framework in parameter selection and estimation and in the key task of GDP forecasting.

REFERENCES


Key words and phrases. Mixed frequency regression, nested lasso, lag selection, spike-and-slab priors.
Regression is the most widely used modeling tool in statistics. Quantile regression offers a strategy for enhancing the regression picture beyond customary mean regression. With time-series data, we move to quantile autoregression and, finally, with spatially referenced time series, we move to space-time quantile regression. Here, we are concerned with the spatiotemporal evolution of daily maximum temperature, particularly with regard to extreme heat. Our motivating data set is 60 years of daily summer maximum temperature data over Aragón in Spain. Hence, we work with time on two scales—days within summer season across years—collected at geocoded station locations. For a specified quantile, we fit a very flexible, mixed-effects autoregressive model, introducing four spatial processes. We work with asymmetric Laplace errors to take advantage of the available conditional Gaussian representation for these distributions. Further, while the autoregressive model yields conditional quantiles, we demonstrate how to extract marginal quantiles with the asymmetric Laplace specification. Thus, we are able to interpolate quantiles for any days within years across our study region.

REFERENCES


Key words and phrases. Asymmetric Laplace distribution, Gaussian process, hierarchical model, marginal quantile, Markov chain Monte Carlo, seasonal time series.


A DYNAMIC SCREENING ALGORITHM FOR HIERARCHICAL BINARY MARKETING DATA

BY YIMEI FAN1,a, YUAN LIAO2,b, ILYA O. RYZHOV3,c AND KUNPENG ZHANG3,d

1Department of Mathematics, University of Maryland, aymfan@math.umd.edu

2Department of Economics, Rutgers University, byuan.liao@rutgers.edu

3Robert H. Smith School of Business, University of Maryland, ciryzhov@umd.edu, dkpzhang@umd.edu

In many applications of business and marketing analytics, predictive models are fit using hierarchically structured data: common characteristics of products, customers, or web pages are represented as categorical variables, and each category can be split up into multiple subcategories at a lower level of the hierarchy. The model may thus contain hundreds of thousands of binary variables, necessitating the use of variable selection to screen out large numbers of irrelevant or insignificant features. We propose a new dynamic screening method, based on the distance correlation criterion, designed for hierarchical binary data. Our method can screen out large parts of the hierarchy at the higher levels, avoiding the need to explore many lower-level features and greatly reducing the computational cost of screening. The practical potential of the method is demonstrated in a case application on user-brand interaction data from Facebook.

REFERENCES


Key words and phrases. Sure screening, hierarchical data, distance correlation.


PENALIZED ESTIMATING EQUATIONS FOR GENERALIZED LINEAR MODELS WITH MULTIPLE IMPUTATION

BY YANG LI\textsuperscript{1,a}, HAOYU YANG\textsuperscript{1,b}, HAOCHEN YU\textsuperscript{1,c}, HANWEN HUANG\textsuperscript{2,d} AND YE SHEN\textsuperscript{2,e}

\textsuperscript{1}Center for Applied Statistics and School of Statistics, Renmin University of China, \textsuperscript{a}yang.li@ruc.edu.cn, \textsuperscript{b}haoyuyang@ruc.edu.cn, \textsuperscript{c}haocheny@ruc.edu.cn
\textsuperscript{2}Department of Epidemiology and Biostatistics, University of Georgia, \textsuperscript{d}huanghw@uga.edu, \textsuperscript{e}yeshen@uga.edu

Missing values among variables present a challenge in variable selection in the generalized linear model. Common strategies that delete observations with missing information may cause serious information loss. Multiple imputation has been widely used in recent years because it provides unbiased statistical results given a correctly specified imputation model and considers the uncertainty of the missing data. However, variable selection methods in the generalized linear model with multiply-imputed data have not yet been studied widely. In this study, we introduce penalized estimating equations for generalized linear models with multiple imputation (PEE–MI), which incorporates the correlation of multiple imputed observations into the objective function. The theoretical performance of the proposed PEE–MI depends on the penalized function adopted. We use the adaptive least absolute shrinkage and selection operator (adaptive LASSO) as an illustrating example. Simulations show that PEE–MI outperforms the alternatives. The proposed method is shown to select variables with clinical relevance when applied to a database of laboratory-diagnosed A/H7N9 patients in the Zhejiang province, China.

REFERENCES


Key words and phrases. Missing data, multiple imputation, penalized estimating equations, variable selection.
SUBBOTIN GRAPHICAL MODELS FOR EXTREME VALUE DEPENDENCIES WITH APPLICATIONS TO FUNCTIONAL NEURONAL CONNECTIVITY

BY ANDERSEN CHANG1,a AND GENEVERA I. ALLEN2,b

1Department of Statistics, Rice University, aatc7@rice.edu
2Department of Electrical and Computer Engineering, Rice University, bgallen@rice.edu

With modern calcium imaging technology, activities of thousands of neurons can be recorded in vivo. These experiments can potentially provide new insights into intrinsic functional neuronal connectivity, defined as contemporaneous correlations between neuronal activities. As a common tool for estimating conditional dependencies in high-dimensional settings, graphical models are a natural choice for estimating functional connectivity networks. However, raw neuronal activity data presents a unique challenge: the relevant information in the data lies in rare extreme value observations that indicate neuronal firing rather than in the observations near the mean. Existing graphical modeling techniques for extreme values rely on binning or thresholding observations which may not be appropriate for calcium imaging data.

In this paper we develop a novel class of graphical models, called the Subbotin graphical model, which finds sparse conditional dependency structures with respect to the extreme value observations without requiring data preprocessing. We first derive the form of the Subbotin graphical model and show the conditions under which it is normalizable. We then study the empirical performance of the Subbotin graphical model and compare it to existing extreme value graphical modeling techniques and functional connectivity models from neuroscience through several simulation studies as well as a real-world calcium imaging data example.

REFERENCES


Key words and phrases. Calcium imaging, exponential family graphical models, extreme values, generalized normal distribution, graphical models, Subbotin distribution.


DOUBLY-ONLINE CHANGEPONT DETECTION FOR MONITORING HEALTH STATUS DURING SPORTS ACTIVITIES

BY MATTIA STIVAL, MAURO BERNARDI AND PETROS DELLAPORTAS

We provide an online framework for analyzing data recorded by smart watches during running activities. In particular, we focus on identifying variations in the behavior of one or more measurements caused by changes in physical condition, such as physical discomfort, periods of prolonged detraining, or even the malfunction of measuring devices. Our framework considers data as a sequence of running activities represented by multivariate time series of physical and biometric data. We combine classical changepoint detection models with an unknown number of components with Gaussian state space models to detect distributional changes between a sequence of activities. The model considers multiple sources of dependence due to the sequential nature of subsequent activities, the autocorrelation structure within each activity, and the contemporaneous dependence between different variables. We provide an online expectation-maximization (EM) algorithm involving a sequential Monte Carlo (SMC) approximation of changepoint predicted probabilities. As a byproduct of our model assumptions, our proposed approach processes sequences of multivariate time series in a doubly-online framework. While classical changepoint models detect changes between subsequent activities, the state space framework, coupled with the online EM algorithm, provides the additional benefit of estimating the real-time probability that a current activity is a changepoint.

REFERENCES


Key words and phrases. Real-time health monitoring, smart watches, online expectation maximization, sequential Monte Carlo.


The SNP-set analysis is a powerful tool for dissecting the genetics of complex human diseases. There are three fundamental genetic association approaches to SNR-set analysis: the marginal model fitting approach, the joint model fitting approach, and the decorrelation approach. A problem of primary interest is how these approaches compare with each other. To address this problem, we develop a theoretical platform to compare the signal-to-noise ratio (SNR) of these approaches under the generalized linear model. We elaborate how causal genetic effects give rise to statistically detectable association signals and show that, when causal effects spread over blocks of strong linkage disequilibrium (LD), the SNR of the marginal model fitting is usually higher than that of the decorrelation approach which, in turn, is higher than that of the unbiased joint model fitting approach. We also scrutinize dense effects and LDs by a bivariate model and extensive simulations using the 1000 Genome Project data. Last, we compare the statistical power of two generic types of SNP-set tests (summation-based and supremum-based) by simulations and an osteoporosis study using large data from UK Biobank. Our results help develop powerful tools for SNP-set analysis and understand the signal detection problem in the presence of colored noise.

REFERENCES


Key words and phrases. SNP-set analysis, causal genetic effect, linkage disequilibrium, signal-noise ratio, global hypothesis test, osteoporosis.


EVALUATING THE USE OF GENERALIZED DYNAMIC WEIGHTED ORDINARY LEAST SQUARES FOR INDIVIDUALIZED HIV TREATMENT STRATEGIES

BY LARRY DONG¹,a, ERICA E. M. MOODIE¹,b, LAURA VILLAIN²,c AND RODOLPHE THIÉBAUT²,d

¹Department of Epidemiology, Biostatistics and Occupational Health, McGill University, a larry.dong@mail.mcgill.ca, b erica.moodie@mcgill.ca
²University of Bordeaux, INSERM U1219 Bordeaux Population Health, Inria SISTM, c laura.villain@u-bordeaux.fr, d rodolphe.thiebaut@u-bordeaux.fr

A dynamic treatment regimes (DTR) represents a statistical paradigm in precision medicine which aims to optimize patient outcomes by individualizing treatments. At its simplest, a DTR may require only a single decision to be made; this special case is called an individualized treatment rule (ITR) and is often used to maximize short-term rewards. Generalized dynamic weighted ordinary least squares (G-dWOLS), a DTR estimation method that offers theoretical advantages such as double robustness of parameter estimators in the decision rules, has been recently extended to accommodate categorical treatments. In this work G-dWOLS is applied to longitudinal data to estimate an optimal ITR. This novel method is demonstrated in simulations and is then applied to a population affected by HIV, whereby an ITR for the administration of Interleukin 7 (IL-7) is devised to maximize the duration where the CD4 load is above a healthy threshold (500 cells/μL) while preventing the administration of unnecessary injections.

REFERENCES


Key words and phrases. Dynamic treatment regime, adaptive treatment strategy, precision medicine, individualized treatment rule, longitudinal data, HIV.
IMPUTATION SCORES

BY JEFFREY NÄF¹, META-LINA SPOHN², LORIS MICHEL³ AND NICOLAI MEINSHAUSEN⁴

Seminar for Statistics, ETH Zürich, ¹naef@stat.math.ethz.ch, ²spohn@stat.math.ethz.ch, ³loris.michel@gmail.com, ⁴meinshausen@stat.math.ethz.ch

Given the prevalence of missing data in modern statistical research, a broad range of methods is available for any given imputation task. How does one choose the “best” imputation method in a given application? The standard approach is to select some observations, set their status to missing, and compare prediction accuracy of the methods under consideration of these observations. Besides having to somewhat artificially mask observations, a shortcoming of this approach is that imputations based on the conditional mean will rank highest if predictive accuracy is measured with quadratic loss. In contrast, we want to rank highest an imputation that can sample from the true conditional distributions. In this paper we develop a framework called “Imputation Scores” (I-Scores) for assessing missing value imputations. We provide a specific I-Score, based on density ratios and projections, that is applicable to discrete and continuous data. It does not require to mask additional observations for evaluations and is also applicable if there are no complete observations. The population version is shown to be proper in the sense that the highest rank is assigned to an imputation method that samples from the correct conditional distribution. The propriety is shown under the missing completely at random (MCAR) assumption but is also shown to be valid under missing at random (MAR) with slightly more restrictive assumptions. We show empirically on a range of data sets and imputation methods that our score consistently ranks true data high(est) and is able to avoid pitfalls usually associated with performance measures such as RMSE. Finally, we provide the R-package Iscores available on CRAN with an implementation of our method.

REFERENCES


Key words and phrases. Ranking, random projections, tree ensembles, random forest, KL-divergence.


AN EFFICIENT DOUBLY-ROBUST IMPUTATION FRAMEWORK FOR LONGITUDINAL DROPOUT, WITH AN APPLICATION TO AN ALZHEIMER’S CLINICAL TRIAL

BY YUQI QIU1,a AND KAREN MESSER2,b

1School of Statistics, East China Normal University, a yqqiu@fem.ecnu.edu.cn
2Division of Biostatistics and Bioinformatics, University of California San Diego, b kmesser@health.ucsd.edu

We develop a novel doubly-robust (DR) imputation framework for longitudinal studies with monotone dropout, motivated by the informative dropout that is common in FDA-regulated trials for Alzheimer’s disease. In this approach the missing data are first imputed using a doubly-robust augmented inverse probability weighting (AIPW) estimator; then the imputed completed data are substituted into a full-data estimating equation, and the estimate is obtained using standard software. The imputed completed data may be inspected and compared to the observed data, and standard model diagnostics are available. The same imputed completed data can be used for several different estimands, such as subgroup analyses in a clinical trial, allowing for reduced computation and increased consistency across analyses. We present two specific DR imputation estimators, AIPW-I and AIPW-S, study their theoretical properties, and investigate their performance by simulation. AIPW-S has substantially reduced computational burden, compared to many other DR estimators, at the cost of some loss of efficiency and the requirement of stronger assumptions. Simulation studies support the theoretical properties and good performance of the DR imputation framework. Importantly, we demonstrate their ability to address time-varying covariates, such as a time by treatment interaction. We illustrate using data from a large randomized Phase III trial, investigating the effect of donepezil in Alzheimer’s disease, from the Alzheimer’s Disease Cooperative Study (ADCS) group.

REFERENCES


Key words and phrases. Doubly-robust estimator, monotone dropouts, imputation methods, Alzheimer’s disease, randomized trials.


A BAYESIAN GROWTH MIXTURE MODEL FOR COMPLEX SURVEY DATA: CLUSTERING POSTDISASTER PTSD TRAJECTORIES

BY REBECCA ANTHOPOLOS¹,a, QIXUAN CHEN²,b, JOSEPH SEDRANSK³,c, MARY THOMPSON⁴,d, GANG MENG⁵,e AND SANDRO GALEA⁶,f

¹Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine, abecca.anthopolos@nyulangone.org
²Department of Biostatistics, Columbia University Mailman School of Public Health, bqc2138@cumc.columbia.edu
³Joint Program in Survey Methodology, University of Maryland, cjxs123@cwru.edu
⁴Department of Statistics and Actuarial Science, University of Waterloo, dmethompson@uwaterloo.ca
⁵Department of Psychology, University of Waterloo, egmeng@uwaterloo.ca
⁶School of Public Health, Boston University, fsgalea@bu.edu

Research on growth mixture models (GMMs) for analyzing data from a complex sample survey is sparse. Existing methods use pseudo-likelihood in which survey weights are incorporated into the likelihood function, with variance estimated via linearization or resampling techniques. Despite popularity of the pseudo-likelihood approach, weighted estimation introduces the risk of efficiency loss. In this paper we propose a Bayesian GMM for complex survey data in which sample design features, such as stratification, clustering, and unequal probability of selection, are incorporated as covariates or hierarchical variance components. The Bayesian GMM can yield a reduction in bias in the estimation of regression coefficients when design features are associated with survey outcomes, and can lead to more efficient estimates than the pseudo-likelihood estimators when the design is noninformative. We develop an efficient Gibbs sampler that includes only closed-form full conditional distributions for model fitting. We present the results of a careful analysis of data from the Galveston Bay Recovery Study (GBRS) which used a stratified multi-stage cluster sample design. Using our proposed Bayesian GMM, we characterize longitudinal trajectories of post-traumatic stress disorder (PTSD) among residents of southeastern Texas in the aftermath of Hurricane Ike. We identify four clinically meaningful PTSD trajectory subgroups and characterize risk factors associated with subgroup membership. In the absence of existing software that can be used to implement our proposed Bayesian GMM for complex survey data, we built the R package Bsvygmm for model fitting, selection, and checking which can be downloaded from https://github.com/anthopolos/Bsvygmm.

REFERENCES


Key words and phrases. Complex survey sample, Gibbs sampling, growth mixture model, post-traumatic stress disorder, spatial modeling.


ESTIMATING HIV EPIDEMICS FOR SUBNATIONAL AREAS

BY LE BAO1,a, XIAOYUE NIU1,b, MARY MAHY2,c AND PETER D. GHYS2,d

1Department of Statistics, Penn State University, a lebao@psu.edu, b xiaoyue@psu.edu
2Strategic Information and Evaluation Department, UNAIDS, c mahym@unaids.org, d ghysp@unaids.org

As the global HIV pandemic enters its fifth decade, increasing numbers of countries use routine HIV testing among pregnant women to monitor their epidemics, allowing governments to look into the epidemics at a finer scale, for example, at subnational levels. Currently, the epidemic model that describes the dynamics of the spread of HIV consists of a set of differential equations and is applied independently to each subnational area. However, the availability of the data varies widely which leads to biased and unreliable estimates for areas with very few data points. We propose to overcome this issue by introducing dependence in the parameters across areas. The proposed method better reconstructs the epidemic trajectories than the independent model as shown in multiple countries in Sub-Saharan Africa. We also offer an approximate method for parameter estimation that is much less computationally burdensome than direct parameter estimation. Compared to direct parameter estimation from the dependent model, the approximate method provides competitive parameter estimation in simulations and the application of HIV subepidemic estimation.

REFERENCES


Key words and phrases. HIV epidemics, correlated model, dynamic system, importance sampling.


UNAIDS (2013). Location, Location: Connecting People Faster to HIV Services. UNAIDS, Geneva, Switzerland.


Massively multiplayer online role-playing games (MMORPGs) offer a unique blend of a personalized gaming experience and a platform for forging social connections. Managers of these digital products rely on predictions of key player responses, such as playing time and purchase propensity, to design timely interventions for promoting, engaging and monetizing their playing base. However, the longitudinal data associated with these MMORPGs not only exhibit a large set of potential predictors to choose from but often present several other distinctive characteristics that pose significant challenges in developing flexible statistical algorithms that can generate efficient predictions of future player activities. For instance, the existence of virtual communities or “guilds” in these games complicate prediction since players who are part of the same guild have correlated behaviors and the guilds themselves evolve over time and thus have a dynamic effect on the future playing behavior of its members. In this paper we develop a crossed random effects joint modeling (CREJM) framework for analyzing correlated player responses in MMORPGs. Contrary to existing methods that assume player independence, CREJM is flexible enough to incorporate both player dependence as well as time-varying guild effects on the future playing behavior of the guild members. On a large-scale data from a popular MMORPG, CREJM conducts simultaneous selection of fixed and random effects in high-dimensional penalized multivariate mixed models. We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency. Besides providing superior predictions of daily playing time and purchase propensity over competing methods, CREJM also predicts player correlations within each guild which are valuable for optimizing future promotional and reward policies for these virtual communities.


STRUCTURE LEARNING FOR ZERO-INFLATED COUNTS WITH AN APPLICATION TO SINGLE-CELL RNA SEQUENCING DATA

BY THI KIM HUE NGUYEN¹,a, KOEN VAN DEN BERGE²,c, MONICA CHIOGNA³,d AND DAVIDE RISSO¹,b

¹Department of Statistical Sciences, University of Padova, a,nguyen@stat.unipd.it, b,davide.risso@unipd.it
²Department of Applied Mathematics, Computer Science and Statistics, Ghent University, c,koenvdberge@berkeley.edu
³Department of Statistical Sciences, University of Bologna, d,monica.chiogna2@unibo.it

The problem of estimating the structure of a graph from observed data is of growing interest in the context of high-throughput genomic data and single-cell RNA sequencing in particular. These, however, are challenging applications, since the data consist of high-dimensional counts with high variance and overabundance of zeros. Here we present a general framework for learning the structure of a graph from single-cell RNA-seq data, based on the zero-inflated negative binomial distribution. We demonstrate with simulations that our approach is able to retrieve the structure of a graph in a variety of settings, and we show the utility of the approach on real data.

REFERENCES


Key words and phrases. Structure learning, graphical models, zero-inflated counts, single-cell RNA-seq.


BAYESIAN INFERENCE AND DYNAMIC PREDICTION FOR MULTIVARIATE LONGITUDINAL AND SURVIVAL DATA

BY HAOTIAN ZOU¹,a, DONGLIN ZENG¹,b, LUO XIAO²,c AND SHENG LUO³,d

¹Department of Biostatistics, University of North Carolina at Chapel Hill, ²haotian@live.unc.edu, bdzeng@email.unc.edu
²Department of Statistics, North Carolina State University, c lxiao5@ncsu.edu
³Department of Biostatistics and Bioinformatics, Duke University, dsheng.luo@duke.edu

Alzheimer’s disease (AD) is a complex neurological disorder impairing multiple domains such as cognition and daily functions. To better understand the disease and its progression, many AD research studies collect multiple longitudinal outcomes that are strongly predictive of the onset of AD dementia. We propose a joint model based on a multivariate functional mixed model framework (referred to as MFMM-JM) that simultaneously models the multiple longitudinal outcomes and the time to dementia onset. We develop six functional forms to fully investigate the complex association between longitudinal outcomes and dementia onset. Moreover, we use the Bayesian methods for statistical inference and develop a dynamic prediction framework that provides accurate personalized predictions of disease progressions based on new subject-specific data. We apply the proposed MFMM-JM to two large ongoing AD studies, the Alzheimer’s Disease Neuroimaging Initiative (ADNI) and National Alzheimer’s Coordinating Center (NACC), and identify the functional forms with the best predictive performance. Our method is also validated by extensive simulation studies with five settings.

REFERENCES


Key words and phrases. Alzheimer’s disease, multivariate longitudinal data, functional mixed model, joint model, Bayesian method, dynamic prediction.
ESTIMATING GARCH(1, 1) IN THE PRESENCE OF MISSING DATA

BY DAMIEN C. H. WEE\textsuperscript{a}, FENG CHEN\textsuperscript{b} AND WILLIAM T. M. DUNSMUIR\textsuperscript{c}

School of Mathematics and Statistics, UNSW Sydney, \textsuperscript{adchwee@gmail.com}, \textsuperscript{bfeng.chen@unsw.edu.au}, \textsuperscript{cw.dunsmuir@unsw.edu.au}

Maximum likelihood estimation of the famous GARCH(1, 1) model is generally straightforward, given the full observation series. However, when some observations are missing, the marginal likelihood of the observed data is intractable in most cases of interest, also intractable is the likelihood from temporally aggregated data. For both these problems, we propose to approximate the intractable likelihoods through sequential Monte Carlo (SMC). The SMC approximation is done in a smooth manner so that the resulting approximate likelihoods can be numerically optimized to obtain parameter estimates. In the case with data aggregation, the use of SMC is made possible by a novel state space representation of the aggregated GARCH series. Through extensive simulation experiments, the proposed method is found to be computationally feasible and produce more accurate estimators of the model parameters compared with other recently published methods, especially in the case with aggregated data. In addition, the Hessian matrix of the minus logarithm of the approximate likelihood can be inverted to produce fairly accurate standard error estimates. The proposed methodology is applied to the analysis of time series data on several exchange-traded funds on the Australian Stock Exchange with missing prices, due to interruptions such as scheduled trading holidays.

REFERENCES


Key words and phrases. Hidden Markov model, incomplete data, intractable likelihood, sequential Monte Carlo, volatility modelling.


**SNIP: AN ADAPTATION OF SORTED NEIGHBORHOOD METHODS FOR DEDUPLICATING PEDIGREE DATA**

**By Theodore Huang**, **Matthew Ploenzke** and **Danielle Braun**

Department of Biostatistics, Harvard T.H. Chan School of Public Health, *thag@ds.dfci.harvard.edu*,

**thuang@ds.dfci.harvard.edu**, *matthew.ploenzke@gmail.com*, *dbraun@mail.harvard.edu*

Pedigree data contain family history information that is used to analyze hereditary diseases. These clinical data sets may contain duplicate records due to the same family visiting a clinic multiple times or a clinician entering multiple versions of the family for testing purposes. Inferences drawn from the data or using them for training or validation without removing the duplicates could lead to invalid conclusions, and hence identifying the duplicates is essential. Since family structures can be complex, direct application of existing deduplication algorithms may not be straightforward. We first motivate the importance of deduplication by examining the impact of pedigree duplicates on model performance when training and validating a familial risk prediction model. We then introduce an unsupervised algorithm, which we call SNIP (Sorted Neighorhood for Pedigrees), that builds on the sorted neighborhood method to find efficiently and to classify pair comparisons by leveraging the inherent hierarchical nature of the pedigrees. We conduct a simulation study to assess the performance of the algorithm and find parameter configurations where the algorithm is able to accurately detect the duplicates. We then apply the method to data from the Risk Service, which includes over 300,000 pedigrees at high risk of hereditary cancers, and uncover large clusters of potential duplicate families. After removing 104,520 pedigrees (33% of original data), the resulting Risk Service data set can now be used for future analysis, training, and validation. The algorithm is available as an R package snipR at [https://github.com/bayesmendel/snipR](https://github.com/bayesmendel/snipR).

**REFERENCES**


*Key words and phrases.* Deduplication, pedigree data, entity matching, sorted neighborhood.
A HORSESHOE MIXTURE MODEL FOR BAYESIAN SCREENING WITH AN APPLICATION TO LIGHT SHEET FLUORESCENCE MICROSCOPY IN BRAIN IMAGING

BY FRANCESCO DENTI\textsuperscript{1,a}, RICARDO AZEVEDO\textsuperscript{2,b}, CHELSIE LO\textsuperscript{2,c}, DAMIAN G. WHEELER\textsuperscript{3,e}, SUNIL P. GANDHI\textsuperscript{2,d}, MICHELE GUINDANI\textsuperscript{4,f} AND BABAK SHAHBABA\textsuperscript{5,g}

\textsuperscript{1}Department of Statistics, Università Cattolica del Sacro Cuore, \textsuperscript{2}Department of Neurobiology and Behavior, University of California, Irvine, \textsuperscript{3}Translucence Biosystems, Inc, \textsuperscript{4}Department of Biostatistics, University of California, Los Angeles, \textsuperscript{5}Department of Statistics, University of California, Irvine.

In this paper we focus on identifying differentially activated brain regions using a light sheet fluorescence microscopy—a recently developed technique for whole-brain imaging. Most existing statistical methods solve this problem by partitioning the brain regions into two classes: significantly and nonsignificantly activated. However, for the brain imaging problem at the center of our study, such binary grouping may provide overly simplistic discoveries by filtering out weak but important signals that are typically adulterated by the noise present in the data. To overcome this limitation, we introduce a new Bayesian approach that allows classifying the brain regions into several tiers with varying degrees of relevance. Our approach is based on a combination of shrinkage priors, widely used in regression and multiple hypothesis testing problems, and mixture models, commonly used in model-based clustering. In contrast to the existing regularizing prior distributions, which use either the spike-and-slab prior or continuous scale mixtures, our class of priors is based on a discrete mixture of continuous scale mixtures and devises a cluster shrinkage version of the horseshoe prior. As a result, our approach provides a more general setting for Bayesian sparse estimation, drastically reduces the number of shrinkage parameters needed, and creates a framework for sharing information across units of interest. We show that this approach leads to more biologically meaningful and interpretable results in our brain imaging problem, since it allows the discrimination between active and inactive regions, while at the same time ranking the discoveries into clusters representing tiers of similar importance.

REFERENCES


Key words and phrases. Bayesian inference, variable selection, mixture models, neuroscience.


WASSERSTEIN, R. L., SCHIRM, A. L. and LAZAR, N. A. (2019). Moving to a world beyond "p < 0.05". *Amer. Statist.* **73** 1–19.
USING PREDICTABILITY TO IMPROVE MATCHING OF URBAN LOCATIONS IN PHILADELPHIA

BY COLMAN HUMPHREY\textsuperscript{a}, RYAN GROSS\textsuperscript{b}, DYLAN S. SMALL\textsuperscript{c} AND SHANE T. JENSEN\textsuperscript{d}

Department of Statistics, The Wharton School, University of Pennsylvania, \textsuperscript{a}colmanhumphrey@gmail.com, \textsuperscript{b}rzgross@wharton.upenn.edu, \textsuperscript{c}dsmall@wharton.upenn.edu, \textsuperscript{d}stjensen@wharton.upenn.edu

Motivated by theories in urban planning and criminology, we use high-resolution data to investigate the relationship between crime and the built environment in the City of Philadelphia. We develop a novel and flexible matching framework that uses the predictability of the treatment variable within matched pairs to empirically inform both the differential weighting of covariates in the matching as well as the selection of the number of matched pairs to create. We use this matching framework for a series of comparisons, each involving matched pairs of Philadelphia intersections that are highly similar on a set of covariates but restricted to differ on a single aspect of the built environment. Our predictability-based matching framework includes data-driven decisions about differential weighting of covariates and the number of matched pairs to create, which is beneficial in our setting as our urban comparisons involve a large number of potential intersections and a large set of covariates to be balanced. In these comparisons we find substantial heterogeneity in the relationships between crime and different aspects of the built environment as well as some empirical support for historical theories.

REFERENCES


\textit{Key words and phrases}. Matching, predictability, urban data, crime.
A SEMIPARAMETRIC PROMOTION TIME CURE MODEL WITH SUPPORT VECTOR MACHINE

BY SUVRA PAL\textsuperscript{a} AND WISDOM ASELISEWINE\textsuperscript{b}

Department of Mathematics, University of Texas at Arlington, \textsuperscript{a}suvra.pal@uta.edu, \textsuperscript{b}wxa5616@mavs.uta.edu

The promotion time cure rate model (PCM) is an extensively studied model for the analysis of time-to-event data in the presence of a cured subgroup. There are several strategies proposed in the literature to model the latency part of PCM. However, there aren’t many strategies proposed to investigate the effects of covariates on the incidence part of PCM. In this regard most existing studies assume the boundary separating the cured and noncured subjects with respect to the covariates to be linear. As such, they can only capture simple effects of the covariates on the cured/noncured probability. In this manuscript we propose a new promotion time cure model that uses the support vector machine (SVM) to model the incidence part. The proposed model inherits the features of the SVM and provides flexibility in capturing nonlinearity in the data. To the best of our knowledge, this is the first work that integrates the SVM with PCM model. For the estimation of model parameters, we develop an expectation maximization algorithm where we make use of the sequential minimal optimization technique together with the Platt scaling method to obtain the posterior probabilities of cured/uncured. A detailed simulation study shows that the proposed model outperforms the existing logistic regression-based PCM model as well as the spline regression-based PCM model, which is also known to capture nonlinearity in the data. This is true in terms of bias and mean square error of different quantities of interest and also in terms of predictive and classification accuracies of cure. Finally, we illustrate the applicability and superiority of our model using the data from a study on leukemia patients who went through bone marrow transplantation.

REFERENCES


Key words and phrases. Machine learning, multiple imputation, sequential minimal optimization, EM algorithm, ROC.


CORRIGENDUM
MODELING BIOMARKER RATIOS WITH GAMMA DISTRIBUTED COMPONENTS

BY MATTHIAS SCHMID

Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, matthias.schmid@imbie.uni-bonn.de

REFERENCES

Exponential Families in Theory and Practice

Bradley Efron, Stanford University

During the past half-century, exponential families have attained a position at the center of parametric statistical inference. Theoretical advances have been matched, and more than matched, in the world of applications, where logistic regression by itself has become the go-to methodology in medical statistics, computer-based prediction algorithms, and the social sciences. This book is based on a one-semester graduate course for first year Ph.D. and advanced master’s students. After presenting the basic structure of univariate and multivariate exponential families, their application to generalized linear models including logistic and Poisson regression is described in detail, emphasizing geometrical ideas, computational practice, and the analogy with ordinary linear regression. Connections are made with a variety of current statistical methodologies: missing data, survival analysis and proportional hazards, false discovery rates, bootstrapping, and empirical Bayes analysis. The book connects exponential family theory with its applications in a way that doesn’t require advanced mathematical preparation.