Function-on-scalar quantile regression with application to mass spectrometry proteomics data ........................................... YUSHA LIU, MENG LI AND JEFFREY S. MORRIS 521
Compression of climate simulations with a nonstationary global SpatioTemporal SPDE model ............................... GEIR-ARNE FUGLSTAD AND STEFANO CASTRUCCIO 542
An efficient and computationally robust statistical method for analyzing case-control mother–offspring pair genetic association studies ......................... HONG ZHANG, BHARAMRUKHERJEE, VICTORIA ARTHUR, GANG HU, HAGIT HOCHEHNER AND JINBO CHEN 560
Baseline drift estimation for air quality data using quantile trend filtering .................................................... HALLEY L. BRANTLEY, JOSPEH GUINNESS AND ERIC C. CHI 585
A fast particle-based approach for calibrating a 3-D model of the Antarctic ice sheet ............................................. BEN SEIYON LEE, MURALI HAKAN, ROBERT W. FULLER, DAVID POLLARD AND KLAUS KELLER 605
Torus graphs for multivariate phase coupling analysis .................................................. NATALIE KLEIN, JOSUE ORELLANA, SCOTT L. BRINCAT, EARL K. MILLER AND ROBERT E. KASS 635
The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments .......................................................... WALTER DEMPSEY, PENG LIAO, SANTOSH KUMAR AND SUSAN A. MURPHY 661
Accounting for uncertainty about past values in probabilistic projections of the total fertility rate for most countries ........... PEIRAN LIU AND ADRIAN E. RAFTERY 685
Estimation of dyadic characteristics of family networks using sample survey data .................................................. CHRIS SKINNER AND FIONA STEELE 706
Sequential importance sampling for multiresolution Kingman–Tajima coalescent counting ....................................... LORENZO CAPPello AND JULIA A. PALACIOS 727
Multiview cluster aggregation and splitting, with an application to multimomic breast cancer data .......................................................... ANTOINE GODICHON-BAGGIONI, CATHY MAUGIS-RABUSSEAU AND ANDREA RAU 752
Semiparametric Bayesian Markov analysis of personalized benefit–risk assessment .................................................. DONGYAN YAN, SUBHARUP GUHA, CHUL AHN AND RAM TIWARI 768
Estimation and inference in metabolomics with nonrandom missing data and latent factors ........................................... CHRIS MCKENNAN, CAROLE OBER AND DAN NICOLAE 789
Bayesian variable selection for survival data using inverse moment priors .......................................................... AMIR NIKOOINEJAD, WENYI WANG AND VALEN E. JOHNSON 809
Evidence factors in a case-control study with application to the effect of flexible sigmoidoscopy screening on colorectal cancer .......................................................... BIKRAM KARMACK, CHYKE A. DOUBENI AND DYLAN S. SMALL 829
A causal exposure response function with local adjustment for confounding: Estimating health effects of exposure to low levels of ambient fine particulate matter .................................................. GEORGIA PAPADOGEORGOU AND FRANCESCA DOMINICI 850
continued
Articles—Continued from front cover

Focused model selection for linear mixed models with an application to whale ecology
CÉLINE CUNEN, LARS WALLØE AND NILS LID HJORT 872

A global-local approach for detecting hotspots in multiple-response regression
HÉLÈNE RUFFIEUX, ANTHONY C. DAVISON, JÖRG HAGER, JAMIE INSHAW, BENJAMIN P. FAIRFAX, SYLVIA RICHARDSON AND LEONARDO BOTTOLLO 905

Seasonal warranty prediction based on recurrent event data
QIANQIAN SHAN, YILI HONG AND WILLIAM Q. MEEKER 929

Generalized accelerated recurrence time model in the presence of a dependent terminal event ............... BO WEI, ZHUMIN ZHANG, HUICHUAN J. LAI AND LIMIN PENG 956

Early identification of an impending rockslide location via a spatially-aided Gaussian mixture model . . . SHUO ZHOU, HOWARD BONDELL, ANTOINETTE TORDESILLAS, BENJAMIN I. P. RUBINSTEIN AND JAMES BAILEY 977

A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging ............... SUBHADEEP PAUL AND YUGUO CHEN 993

Analyses of preventive care measures with incomplete historical data in electronic medical records: An example from colorectal cancer screening . . . . YINGYE ZHENG, DOUGLAS A. CORLEY, CHYKE DOUBENI, ETHAN HALM, SUSAN M. SHORTREED, WILLIAM E. BARLOW, ANN ZAUBER, TOR DEVIN TOSTESON AND JESSICA CHUBAK 1030

Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples and multiple imputation
MARK J. GIGANTI, PAMELA A. SHAW, GUANHUA CHEN, SALLY S. BEBAWY, MEGAN M. TURNER, TIMOTHY R. STERLING AND BRYAN E. SHEPHERD 1045
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: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

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

Past President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

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

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

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS PUBLICATIONS

The Annals of Statistics. Editors: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

The Annals of Applied Statistics. Editor-In-Chief: Karen Kafadar, Department of Statistics, University of Virginia, Charlottesville, VA 22904-4135, USA

The Annals of Probability. Editor: Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA


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

The IMS Bulletin. Editor: Vlada Limic, UMR 7501 de l’Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Applied Statistics [ISSN 1932-6157 (print); ISSN 1941-7330 (online)], Volume 14, Number 2, June 2020. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, 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, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

Copyright © 2020 by the Institute of Mathematical Statistics
Printed in the United States of America
FUNCTION-ON-SCALAR QUANTILE REGRESSION WITH APPLICATION TO MASS SPECTROMETRY PROTEOMICS DATA

BY YUSHA LIU1,*, MENG LI1,† AND JEFFREY S. MORRIS2

1Department of Statistics, Rice University, *yl95@rice.edu; †meng@rice.edu
2Division of Biostatistics, University of Pennsylvania, Jeffrey.Morris@pennmedicine.upenn.edu

Mass spectrometry proteomics, characterized by spiky, spatially heterogeneous functional data, can be used to identify potential cancer biomarkers. Existing mass spectrometry analyses utilize mean regression to detect spectral regions that are differentially expressed across groups. However, given the interpatient heterogeneity that is a key hallmark of cancer, many biomarkers are only present at aberrant levels for a subset of, not all, cancer samples. Differences in these biomarkers can easily be missed by mean regression but might be more easily detected by quantile-based approaches. Thus, we propose a unified Bayesian framework to perform quantile regression on functional responses. Our approach utilizes an asymmetric Laplace working likelihood, represents the functional coefficients with basis representations which enable borrowing of strength from nearby locations and places a global-local shrinkage prior on the basis coefficients to achieve adaptive regularization. Different types of basis transform and continuous shrinkage priors can be used in our framework. A scalable Gibbs sampler is developed to generate posterior samples that can be used to perform Bayesian estimation and inference while accounting for multiple testing. Our framework performs quantile regression and coefficient regularization in a unified manner, allowing them to inform each other and leading to improvement in performance over competing methods. We also introduce an adjustment procedure to the model to improve its frequentist properties of posterior inference. We apply our model to identify proteomic biomarkers of pancreatic cancer that are differentially expressed for a subset of cancer patients compared to the normal controls which were missed by previous mean-regression based approaches. Supplementary Material for this article is available online.

REFERENCES


Key words and phrases. Bayesian hierarchical model, functional data analysis, functional response regression, global-local shrinkage, proteomic biomarker, quantile regression.


COMPRESSION OF CLIMATE SIMULATIONS WITH A NONSTATIONARY GLOBAL SPATIOTEMPORAL SPDE MODEL

BY GEIR-ARNE FUGLSTAD¹ AND STEFANO CASTRUCCIO²

¹Department of Mathematical Sciences, Norwegian University of Science and Technology, geir-arne.fuglstad@ntnu.no
²Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, scastruc@nd.edu

Modern climate models pose an ever-increasing storage burden to computational facilities, and the upcoming generation of global simulations from the next Intergovernmental Panel on Climate Change will require a substantial share of the budget of research centers worldwide to be allocated just for this task. A statistical model can be used as a means to mitigate the storage burden by providing a stochastic approximation of the climate simulations. Indeed, if a suitably validated statistical model can be formulated to draw realizations whose spatiotemporal structure is similar to that of the original computer simulations, then the estimated parameters are effectively all the information that needs to be stored. In this work we propose a new statistical model defined via a stochastic partial differential equation (SPDE) on the sphere and in evolving time. The model is able to capture nonstationarities across latitudes, longitudes and land/ocean domains for more than 300 million data points while also overcoming the fundamental limitations of current global statistical models available for compression. Once the model is trained, surrogate runs can be instantaneously generated on a laptop by storing just 20 Megabytes of parameters as opposed to more than six Gigabytes of the original ensemble.

REFERENCES


Key words and phrases. Stochastic partial differential equation, space-time model, global model, nonstationary, climate model.


AN EFFICIENT AND COMPUTATIONALLY ROBUST STATISTICAL METHOD FOR ANALYZING CASE-CONTROL MOTHER–OFFSPRING PAIR GENETIC ASSOCIATION STUDIES

BY HONG ZHANG1, BHRAMAR MUKHERJEE2, VICTORIA ARTHUR3,*, GANG HU4, HAGIT HOCHNER5 AND JINBO CHEN3,**

1Department of Statistics and Finance, School of Management, University of Science and Technology of China, zhangh@ustc.edu.cn
2Department of Biostatistics, University of Michigan, bhramar@umich.edu
3Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, *vlynn@penmedicine.upenn.edu; **jinboche@mail.med.upenn.edu
4Chronic Disease Epidemiology Laboratory, Pennington Biomedical Research Center, gang.hu@pbrc.edu
5The Hebrew University of Jerusalem, hagith@ekmd.huji.ac.il

Case-control mother–offspring pair design has been widely adopted for studying early-life and women’s pregnancy health. It allows assessment of pre- and perinatal environmental risk factors as well as both maternal and offspring genetic risk factors. Data arising from this design is routinely analyzed using standard prospective logistic regression. Such data has two unique features: the offspring genotypes are not correlated with maternal environmental risk factors given maternal genotypes, and offspring and maternal genotypes are related through mendelian transmission. In this work, built upon a novel regression model relating maternal genotypes to environmental risk factors, we proposed a novel retrospective likelihood method that effectively utilized the two data features to increase statistical efficiency for detecting maternal and offspring genetic effects. The inference procedure was based on a profile likelihood derived using the Lagrange multiplier method, but we replaced the multipliers with their large sample limits to enable highly efficient and computationally stable estimation. We showed that our proposed estimates of odds ratio association parameters are consistent and asymptotically normally distributed and demonstrated the finite sample performance through extensive simulation studies and application to genetic association studies of birth weight and gestational diabetes mellitus.

REFERENCES


Key words and phrases. Case-control mother–offspring pair design, genetic association, profile likelihood, saddle point problem, retrospective likelihood.


BASELINE DRIFT ESTIMATION FOR AIR QUALITY DATA USING QUANTILE TREND FILTERING

BY HALLEY L. BRANTLEY1,*, JOSEPH GUINNESS2 AND ERIC C. CHI1,†

1Department of Statistics, North Carolina State University, * hlbrantl@ncsu.edu; † eric_chi@ncsu.edu
2Department of Statistics and Data Science, Cornell University, guinness@cornell.edu

We address the problem of estimating smoothly varying baseline trends in time series data. This problem arises in a wide range of fields, including chemistry, macroeconomics and medicine; however, our study is motivated by the analysis of data from low cost air quality sensors. Our methods extend the quantile trend filtering framework to enable the estimation of multiple quantile trends simultaneously while ensuring that the quantiles do not cross. To handle the computational challenge posed by very long time series, we propose a parallelizable alternating direction method of multipliers (ADMM) algorithm. The ADMM algorithm enables the estimation of trends in a piecewise manner, both reducing the computation time and extending the limits of the method to larger data sizes. We also address smoothing parameter selection and propose a modified criterion based on the extended Bayesian information criterion. Through simulation studies and our motivating application to low cost air quality sensor data, we demonstrate that our model provides better quantile trend estimates than existing methods and improves signal classification of low-cost air quality sensor output.

REFERENCES


Key words and phrases. Air quality, nonparametric quantile regression, trend estimation.
A FAST PARTICLE-BASED APPROACH FOR CALIBRATING A 3-D MODEL OF THE ANTARCTIC ICE SHEET

BY BEN SEIYON LEE1,*, MURALI HARAN1,†, ROBERT W. FULLER2,‡, DAVID POLLARD3 AND KLAUS KELLER2,§

1Department of Statistics, Pennsylvania State University, *skl5261@psu.edu; †mharan@stat.psu.edu
2Department of Geosciences, Pennsylvania State University, ‡rwf136@psu.edu; §klaus@psu.edu
3Earth and Environmental Systems Institute, Pennsylvania State University, pollard@essc.psu.edu

We consider the scientifically challenging and policy-relevant task of understanding the past and projecting the future dynamics of the Antarctic ice sheet. The Antarctic ice sheet has shown a highly nonlinear threshold response to past climate forcings. Triggering such a threshold response through anthropogenic greenhouse gas emissions would drive drastic and potentially fast sea level rise with important implications for coastal flood risks. Previous studies have combined information from ice sheet models and observations to calibrate model parameters. These studies have broken important new ground but have either adopted simple ice sheet models or have limited the number of parameters to allow for the use of more complex models. These limitations are largely due to the computational challenges posed by calibration as models become more computationally intensive or when the number of parameters increases.

Here, we propose a method to alleviate this problem: a fast sequential Monte Carlo method that takes advantage of the massive parallelization afforded by modern high-performance computing systems. We use simulated examples to demonstrate how our sample-based approach provides accurate approximations to the posterior distributions of the calibrated parameters. The drastic reduction in computational times enables us to provide new insights into important scientific questions, for example, the impact of Pliocene era data and prior parameter information on sea level projections. These studies would be computationally prohibitive with other computational approaches for calibration such as Markov chain Monte Carlo or emulation-based methods. We also find considerable differences in the distributions of sea level projections when we account for a larger number of uncertain parameters. For example, based on the same ice sheet model and data set, the 99th percentile of the Antarctic ice sheet contribution to sea level rise in 2300 increases from 6.5 m to 13.1 m when we increase the number of calibrated parameters from three to 11. With previous calibration methods, it would be challenging to go beyond five parameters. This work provides an important next step toward improving the uncertainty quantification of complex, computationally intensive and decision-relevant models.

REFERENCES


Key words and phrases. Antarctic ice sheet model, computer model calibration, paleoclimate, sequential Monte Carlo, uncertainty quantification.


Angular measurements are often modeled as circular random variables, where there are natural circular analogues of moments, including correlation. Because a product of circles is a torus, a $d$-dimensional vector of circular random variables lies on a $d$-dimensional torus. For such vectors we present here a class of graphical models, which we call torus graphs, based on the full exponential family with pairwise interactions. The topological distinction between a torus and Euclidean space has several important consequences.

Our development was motivated by the problem of identifying phase coupling among oscillatory signals recorded from multiple electrodes in the brain: oscillatory phases across electrodes might tend to advance or recede together, indicating coordination across brain areas. The data analyzed here consisted of 24 phase angles measured repeatedly across 840 experimental trials (replications) during a memory task, where the electrodes were in 4 distinct brain regions, all known to be active while memories are being stored or retrieved. In realistic numerical simulations, we found that a standard pairwise assessment, known as phase locking value, is unable to describe multivariate phase interactions, but that torus graphs can accurately identify conditional associations. Torus graphs generalize several more restrictive approaches that have appeared in various scientific literatures, and produced intuitive results in the data we analyzed. Torus graphs thus unify multivariate analysis of circular data and present fertile territory for future research.

**REFERENCES**


*Key words and phrases.* Graphical models, circular statistics, network analysis.
https://doi.org/10.1017/CBO9780511564345

https://doi.org/10.1016/j.laa.2014.08.015


https://doi.org/10.1016/j.csda.2006.09.003


THE STRATIFIED MICRO-RANDOMIZED TRIAL DESIGN: SAMPLE SIZE CONSIDERATIONS FOR TESTING NESTED CAusal EFFECTS OF TIME-VARYING TREATMENTS

BY WALTER DEMPSEY1,*, PENG LIAO1,**, SANTOSH KUMAR2 AND SUSAN A. MURPHY3

1Department of Biostatistics, University of Michigan, * wdem@umich.edu; ** pengliao@umich.edu
2Department of Computer Science, University of Memphis, santosh.kumar@memphis.edu
3Department of Statistics, Harvard University, samurphy@fas.harvard.edu

Technological advancements in the field of mobile devices and wearable sensors have helped overcome obstacles in the delivery of care, making it possible to deliver behavioral treatments anytime and anywhere. Here, we discuss our work on the design of a mobile health smoking cessation intervention study with the goal of assessing whether reminders, delivered at times of stress, result in a reduction/prevention of stress in the near-term and whether this effect changes with time in study. Multiple statistical challenges arose in this effort, leading to the development of the 

stratified micro-randomized trial design. In these designs each individual is randomized to treatment repeatedly at times determined by predictions of risk. These risk times may be impacted by prior treatment. We describe the statistical challenges and detail how they can be met.

REFERENCES


Key words and phrases. Sequential randomization, nested causal effects, stratified micro-randomized trials, mobile health, weighted-centered least-squares method.


ACCOUNTING FOR UNCERTAINTY ABOUT PAST VALUES IN PROBABILISTIC PROJECTIONS OF THE TOTAL FERTILITY RATE FOR MOST COUNTRIES

BY PEIRAN LIU* AND ADRIAN E. RAFTERY**

Department of Statistics, University of Washington, *prliu@uw.edu; **raftery@uw.edu

Since the 1940s, population projections have in most cases been produced using the deterministic cohort component method. However, in 2015, for the first time and in a major advance, the United Nations issued official probabilistic population projections for all countries based on Bayesian hierarchical models for total fertility and life expectancy. The estimates of these models and the resulting projections are conditional on the U.N.’s official estimates of past values. However, these past values are themselves uncertain, particularly for the majority of the world’s countries that do not have long-standing high-quality vital registration systems, when they rely on surveys and censuses with their own biases and measurement errors. This paper extends the U.N. model for projecting future total fertility rates to take account of uncertainty about past values. This is done by adding an additional level to the hierarchical model to represent the multiple data sources, in each case estimating their bias and measurement error variance. We assess the method by out-of-sample predictive validation. While the prediction intervals produced by the extant method (which does not account for this source of uncertainty) have somewhat less than nominal coverage, we find that our proposed method achieves closer to nominal coverage. The prediction intervals become wider for countries for which the estimates of past total fertility rates rely heavily on surveys rather than on vital registration data, especially in high fertility countries.

REFERENCES


Key words and phrases. Bayesian hierarchical model, Markov chain Monte Carlo, measurement error, Population projection, total fertility rate, vital registration.


ESTIMATION OF DYADIC CHARACTERISTICS OF FAMILY NETWORKS USING SAMPLE SURVEY DATA

BY CHRIS SKINNER* AND FIONA STEELE†

Department of Statistics, London School of Economics and Political Science, * c.j.skinner@lse.ac.uk; † f.a.steele@lse.ac.uk

We consider the use of sample survey data to estimate dyadic characteristics of family networks, with an application to noncoresident parent–child dyads. We suppose that survey respondents report either from a parent or child perspective about a dyad, depending on their membership of the dyad. We construct separate estimators of common dyadic characteristics using data from both a parent and a child perspective and show how comparisons of these estimators can shed light on data quality issues including differential missingness and reporting error. In our application we find that a simple missingness model explains some striking patterns of discrepancies between the estimators and consider the use of poststratification and a related redefinition of count variables to adjust for these discrepancies. We also develop approaches to combining the separate estimators efficiently to estimate means and frequency distributions within subpopulations.

REFERENCES


**Key words and phrases.** Coresident, network sampling, poststratification, reporting error.


SEQUENTIAL IMPORTANCE SAMPLING FOR MULTIRESOLUTION KINGMAN–TAJIMA COALESCENT COUNTING

BY LORENZO CAPPELLO1,* AND JULIA A. PALACIOS1,2,†

1Department of Statistics, Stanford University
2Department of Biomedical Data Science, Stanford University, *cappello@stanford.edu; †juliapr@stanford.edu

Statistical inference of evolutionary parameters from molecular sequence data relies on coalescent models to account for the shared genealogical ancestry of the samples. However, inferential algorithms do not scale to available data sets. A strategy to improve computational efficiency is to rely on simpler coalescent and mutation models, resulting in smaller hidden state spaces. An estimate of the cardinality of the state space of genealogical trees at different resolutions is essential to decide the best modeling strategy for a given dataset. To our knowledge, there is neither an exact nor approximate method to determine these cardinalities. We propose a sequential importance sampling algorithm to estimate the cardinality of the sample space of genealogical trees under different coalescent resolutions. Our sampling scheme proceeds sequentially across the set of combinatorial constraints imposed by the data which, in this work, are completely linked sequences of DNA at a non-recombining segment. We analyze the cardinality of different genealogical tree spaces on simulations to study the settings that favor coarser resolutions. We apply our method to estimate the cardinality of genealogical tree spaces from mtDNA data from the 1000 genomes and a sample from a Melanesian population at the β-globin locus.

REFERENCES


Key words and phrases. Coalescent, sequential importance sampling, enumeration.


MULTIVIEW CLUSTER AGGREGATION AND SPLITTING, WITH AN APPLICATION TO MULTOMIC BREAST CANCER DATA

BY ANTOINE GODICHON-BAGGIONI¹, CATHY MAUGIS-RABUSSEAU² AND ANDREA RAU³

¹Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, antoine.godichon_baggioni@upmc.fr
²Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, CNRS, INSA Toulouse, cathy.maugis@insa-toulouse.fr
³Université Paris-Saclay, INRAE, AgroParisTech, GABI, andrea.rau@inrae.fr

Multiview data, which represent distinct but related groupings of variables, can be useful for identifying relevant and robust clustering structures among observations. A large number of multiview classification algorithms have been proposed in the fields of computer science and genomics; here, we instead focus on the task of merging or splitting an existing hard or soft cluster partition based on multiview data. This article is specifically motivated by an application involving multomic breast cancer data from The Cancer Genome Atlas, where multiple molecular profiles (gene expression, microRNA expression, methylation and copy number alterations) are used to further subdivide the five currently accepted intrinsic tumor subtypes into distinct subgroups of patients. In addition, we investigate the performance of the proposed multiview splitting and aggregation algorithms, as compared to single- and concatenated-view alternatives, in a set of simulations. The multiview splitting and aggregation algorithms developed here are implemented in the maskmeans R package.

REFERENCES


Key words and phrases. Clustering, multiview, cluster merging and splitting, multiomic data, TCGA.


SEMIPARAMETRIC BAYESIAN MARKOV ANALYSIS OF PERSONALIZED BENEFIT–RISK ASSESSMENT

BY DONGYAN YAN1, SUBHARUP GUHA2, CHUL AHN3, * AND RAM TIWARI3, †

1Discovery & Development Statistics, Eli Lilly and Company, dyyr2@mail.missouri.edu
2Department of Biostatistics, University of Florida, s.guha@ufl.edu
3Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, U.S. Food and Drug Administration, *chul.ahn@fda.hhs.gov; †ram.tiwari@fda.hhs.gov

The development of systematic and structured approaches to assess benefit–risk of medical products is a major challenge for regulatory decision makers. Existing benefit–risk methods depend only on the frequencies of mutually exclusive and exhaustive categories in which the subjects fall, and the responses of individuals are allowed to belong to any of the other categories during their postwithdrawal visits. In this article we introduce a semiparametric Bayesian Markov model (SBMM) that treats the withdrawal category as an absorbing state and analyzes subject-level data for multiple visits, accounting for any within-patient dependencies in the response profiles. A log-odds ratio model is used to model the subject-level effects by assuming a ratio of transition probabilities with respect to a “reference” category. A Dirichlet process is used as a semiparametric model for the subject-level effects to flexibly capture the underlying distributions of the personalized response profiles without making strong parametric assumptions. This also allows the borrowing of strength between the patients and achieves dimension reduction by allocating similar response profiles patterns into an unknown number of latent clusters. We analyze a motivating clinical trial dataset to assess the personalized benefit–risks in each arm and evaluate the aggregated benefits and risks associated with the drug Exalgo.

REFERENCES


Key words and phrases. Clinical trials, Dirichlet process, Exalgo, Gibbs sampling, log-odds ratio model, Metropolis–Hastings, SBMM.


ESTIMATION AND INFERENCE IN METABOLOMICS WITH NONRANDOM MISSING DATA AND LATENT FACTORS

BY CHRIS MCKENNAN¹, CAROLE OBER² AND DAN NICOLAE³

¹Department of Statistics, University of Pittsburgh, chm195@pitt.edu
²Department of Human Genetics, University of Chicago, c-ober@bsd.uchicago.edu
³Department of Statistics, University of Chicago, nicolae@galton.uchicago.edu

High-throughput metabolomics data are fraught with both nonignorable missing observations and unobserved factors that influence a metabolite’s measured concentration, and it is well known that ignoring either of these complications can compromise estimators. However, current methods to analyze these data can only account for the missing data or unobserved factors, but not both. We therefore developed MetabMiss, a statistically rigorous method to account for both nonrandom missing data and latent factors in high-throughput metabolomics data. Our methodology does not require the practitioner specify a likelihood for the missing data, and makes investigating the relationship between the metabolome and tens, or even hundreds, of phenotypes computationally tractable. We demonstrate the fidelity of MetabMiss’s estimates using both simulated and real metabolomics data and prove their asymptotic correctness when the sample size and number of metabolites grows to infinity.

REFERENCES


Key words and phrases. Metabolomics, latent factors, batch variables, generalized method of moments, missing not at random (MNAR).


BAYESIAN VARIABLE SELECTION FOR SURVIVAL DATA USING INVERSE MOMENT PRIORS

BY AMIR NIKOOEIENEJAD¹, WENYI WANG² AND VALEN E. JOHNSON³

¹Department of Global Statistical Sciences, Eli Lilly and Company, nikooienejad_amir@lilly.com
²Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, w wang7@mdanderson.org
³Department of Statistics, Texas A&M University, vjohnson@stat.tamu.edu

Efficient variable selection in high-dimensional cancer genomic studies is critical for discovering genes associated with specific cancer types and for predicting response to treatment. Censored survival data is prevalent in such studies. In this article we introduce a Bayesian variable selection procedure that uses a mixture prior composed of a point mass at zero and an inverse moment prior in conjunction with the partial likelihood defined by the Cox proportional hazard model. The procedure is implemented in the R package BVSNLP, which supports parallel computing and uses a stochastic search method to explore the model space. Bayesian model averaging is used for prediction. The proposed algorithm provides better performance than other variable selection procedures in simulation studies and appears to provide more consistent variable selection when applied to actual genomic datasets.

REFERENCES


Key words and phrases. Bayesian variable selection, nonlocal prior, high-dimensional data, survival data analysis, Cox proportional hazard model, cancer genomics.


EVIDENCE FACTORS IN A CASE-CONTROL STUDY WITH APPLICATION TO THE EFFECT OF FLEXIBLE SIGMOIDOSCOPY SCREENING ON COLORECTAL CANCER

BY BIKRAM KARMAKAR¹, CHYKE A. DOUBENI² AND DYLAN S. SMALL³

¹Department of Statistics, College of Liberal Arts and Sciences, University of Florida, bkarmakar@ufl.edu
²Center for Health Equity and Community Engagement Research, Mayo Clinic, Doubeni.Chyke@mayo.edu
³Department of Statistics, The Wharton School, University of Pennsylvania, dsmall@wharton.upenn.edu

As in any observational study, in a case-control study a primary concern is potential unmeasured confounders. Bias, due to unmeasured confounders, can result in a false discovery of an apparent treatment effect when there is none. Replication of an observational study, which tries to provide multiple analyses of the data where the biases affecting each analysis are thought to be different, is one way to strengthen the evidence from an observational study. Evidence factors allow for internal replication by testing a hypothesis using multiple comparisons in a way that the comparisons yield independent evidence and differ in the sources of potential bias. We construct evidence factors in a case-control study in which there are two types of cases, “narrow” cases which are thought to be potentially more affected by the exposure and “marginal” cases which are thought to have more heterogeneous causes. We develop and study an inference procedure for using such evidence factors and apply it to a study of the effect of sigmoidoscopy screening on colorectal cancer.

REFERENCES


Key words and phrases. Case-control studies, colorectal cancer, evidence factors, observational study, replicability, sigmoidoscopy.


A CAUSAL EXPOSURE RESPONSE FUNCTION WITH LOCAL ADJUSTMENT FOR CONFOUNDING: ESTIMATING HEALTH EFFECTS OF EXPOSURE TO LOW LEVELS OF AMBIENT FINE PARTICULATE MATTER

BY GEORGIA PAPADOGEORGOU¹ AND FRANCESCA DOMINICI²

¹Department of Statistical Science, Duke University, gp118@duke.edu
²Department of Biostatistics, Harvard T.H. Chan School of Public Health, fdominic@hsph.harvard.edu

In the last two decades ambient levels of air pollution have declined substantially. At the same time the Clean Air Act mandates that the National Ambient Air Quality Standards (NAAQS) must be routinely assessed to protect populations based on the latest science. Therefore, researchers should continue to address the following question: is exposure to levels of air pollution below the NAAQS harmful to human health? Furthermore, the contentious nature surrounding environmental regulations urges us to cast this question within a causal inference framework. Several parametric and semiparametric regression approaches have been used to estimate the exposure-response (ER) curve between long-term exposure to ambient air pollution concentrations and health outcomes. However, most of the existing approaches are not formulated within a formal framework for causal inference, adjust for the same set of potential confounders across all levels of exposure and do not account for model uncertainty regarding covariate selection and the shape of the ER.

In this paper we introduce a Bayesian framework for the estimation of a causal ER curve called LERCA (Local Exposure Response Confounding Adjustment), which: (a) allows for different confounders and different strength of confounding at the different exposure levels, and (b) propagates model uncertainty regarding confounders’ selection and the shape of the ER. Importantly, LERCA provides a principled way of assessing the observed covariates’ confounding importance at different exposure levels, providing researchers with important information regarding the set of variables to measure and adjust for in regression models. Using simulation studies, we show that state-of-the-art approaches perform poorly in estimating the ER curve in the presence of local confounding.

LERCA is used to evaluate the relationship between long-term exposure to ambient PM₂.₅, a key regulated pollutant, and cardiovascular hospitalizations for 5,362 zip codes in the continental U.S. and located near a pollution monitoring site, while adjusting for a potentially varying set of confounders across the exposure range. Our data set includes rich health, weather, demographic and pollution information for the years of 2011–2013. The estimated exposure-response curve is increasingly indicating that higher ambient concentrations lead to higher cardiovascular hospitalization rates, and ambient PM₂.₅ was estimated to lead to an increase in cardiovascular hospitalization rates when focusing at the low-exposure range. Our results indicate that there is no threshold for the effect of PM₂.₅ on cardiovascular hospitalizations.

REFERENCES


Key words and phrases. Air pollution, cardiovascular hospitalizations, causal inference, exposure response function, local confounding, low-exposure levels, particulate matter.


VANSTEELANDT, S., BEKAERT, M. and CLAESKENS, G. (2012). On model selection and model mis-
0962280210387717
VODONOS, A., AWAD, Y. A. and SCHWARTZ, J. (2018). The concentration-response between long-term PM$_{2.5}$
envres.2018.06.021
WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment
WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in con-
founder and effect modifier selection when estimating average causal effects in generalized linear models.
exposure to PM2.5 and mortality among older adults in the southeastern US. Epidemiology 28 207–214.
https://doi.org/10.1097/EDE.0000000000000614.
WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information
WEICHENTHAL, S., KULKA, R., LAVIGNE, E., RIJSWIJK, D. V., BRAUER, M., VILLENEUVE, P. J., STIEB, D.,
JOSEPH, L. and BURNETT, R. T. (2017). Biomass burning as a source of ambient fine particulate air pollution
and acute myocardial infarction. Epidemiology 28 329–337. https://doi.org/10.1097/EDE.0000000000000636
861. MR3295746 https://doi.org/10.1111/biom.12203
ZANOBETTI, A. and SCHWARTZ, J. (2007). Particulate air pollution, progression, and survival after myocardial
116 1614–1619. https://doi.org/10.1289/ehp.11449
ZIGLER, C. M. and DOMINICI, F. (2014). Point: Clarifying policy evidence with potential-outcomes thinking-
FOCUSED MODEL SELECTION FOR LINEAR MIXED MODELS WITH AN APPLICATION TO WHALE ECOLOGY

BY CÉLINE CUNEN¹,*, LARS WALLØE² AND NILS LID HJORT¹,†

¹Department of Mathematics, University of Oslo, *cmlcunen@math.uio.no; †nils@math.uio.no
²Institute of Basic Medical Sciences, University of Oslo, lars.walloe@medisin.uio.no

A central point of disagreement, in certain long-standing discussions about a particular whaling dataset in the Scientific Committee of the International Whaling Commission, has directly involved model selection issues for linear mixed effect models. The biological question under discussion is associated with a clearly defined parameter of primary interest, a focus parameter, which makes model selection with the Focused Information Criterion (FIC) more appropriate than other selection methods. Since the existing FIC methodology has not covered the case of linear mixed effects models, this article sets up the required framework and develops the necessary formulae for the relevant FIC. Our new criterion requires the asymptotic distribution of estimators derived for a given candidate linear mixed model but with behaviour examined under a wider linear mixed model. These results, needed here to build our FIC, also have independent interest.

REFERENCES


Key words and phrases. Antarctic minke whale, focused information criterion, FIC plot, linear mixed effects models, model selection.


A GLOBAL-LOCAL APPROACH FOR DETECTING HOTSPOTS IN MULTIPLE-RESPONSE REGRESSION

BY HÉLÈNE RUFFIEUX1, ANTHONY C. DAVISON2, JÖRG HAGER3, JAMIE INSHAW4, BENJAMIN P. FAIRFAX5, SYLVIA RICHARDSON1,6 AND LEONARDO BOTTOLO7,6,1

We tackle modelling and inference for variable selection in regression problems with many predictors and many responses. We focus on detecting hotspots, that is, predictors associated with several responses. Such a task is critical in statistical genetics, as hotspot genetic variants shape the architecture of the genome by controlling the expression of many genes and may initiate decisive functional mechanisms underlying disease endpoints. Existing hierarchical regression approaches designed to model hotspots suffer from two limitations: their discrimination of hotspots is sensitive to the choice of top-level scale parameters for the propensity of predictors to be hotspots, and they do not scale to large predictor and response vectors, for example, of dimensions $10^3$–$10^5$ in genetic applications. We address these shortcomings by introducing a flexible hierarchical regression framework that is tailored to the detection of hotspots and scalable to the above dimensions. Our proposal implements a fully Bayesian model for hotspots based on the horseshoe shrinkage prior. Its global-local formulation shrinks noise globally and, hence, accommodates the highly sparse nature of genetic analyses while being robust to individual signals, thus leaving the effects of hotspots unshrunk. Inference is carried out using a fast variational algorithm coupled with a novel simulated annealing procedure that allows efficient exploration of multimodal distributions.

REFERENCES


Key words and phrases. Annealed variational inference, hierarchical model, horseshoe prior, molecular quantitative trait locus analyses, multiplicity control, normal scale mixture, regulation hotspot, shrinkage, statistical genetics, variable selection.


SEASONAL WARRANTY PREDICTION BASED ON RECURRENT EVENT DATA

BY QIANQIAN SHAN\textsuperscript{1,*}, YILI HONG\textsuperscript{2} AND WILLIAM Q. MEEKER\textsuperscript{1,†}

\textsuperscript{1}Department of Statistics, Iowa State University, \textsuperscript{*}qshan@iastate.edu; \textsuperscript{†}wqmeeker@iastate.edu
\textsuperscript{2}Department of Statistics, Virginia Tech, yilihong@vt.edu

Warranty return data from repairable systems, such as home appliances, lawn mowers, computers and automobiles, result in recurrent event data. The nonhomogeneous Poisson process (NHPP) model is used widely to describe such data. Seasonality in the repair frequencies and other variabilities, however, complicate the modeling of recurrent event data. Not much work has been done to address the seasonality, and this paper provides a general approach for the application of NHPP models with dynamic covariates to predict seasonal warranty returns. The methods presented here, however, can be applied to other applications that result in seasonal recurrent event data. A hierarchical clustering method is used to stratify the population into groups that are more homogeneous than the overall population. The stratification facilitates modeling the recurrent event data with both time-varying and time-constant covariates. We demonstrate and validate the models using warranty claims data for two different types of products. The results show that our approach provides important improvements in the predictive power of monthly events compared with models that do not take the seasonality and covariates into account.

REFERENCES


Key words and phrases. EM algorithm, hierarchical clustering, missing data, NHPP, random effects, seasonal dynamic covariates.
GENERALIZED ACCELERATED RECURRENCE TIME MODEL IN THE PRESENCE OF A DEPENDENT TERMINAL EVENT

BY BO WEI1,*, ZHUMIN ZHANG2,‡, HUICHUAN J. LAI2,§ AND LIMIN PENG1,†

1Department of Biostatistics and Bioinformatics, Emory University, *bo.wei@emory.edu; †lpeng@emory.edu
2Departments of Nutritional Sciences, University of Wisconsin–Madison, ‡zhuminzhang@wisc.edu; §hlai@wisc.edu

Recurrent events are commonly encountered in longitudinal studies. The observation of recurrent events is often stopped by a dependent terminal event in practice. For this data scenario, we propose two sensible adaptations of the generalized accelerated recurrence time (GART) model (J. Amer. Statist. Assoc. 111 (2016) 145–156) to provide useful alternative analyses that can offer physical interpretations while rendering extra flexibility beyond the existing work based on the accelerated failure time model. Our modeling strategies align with the rationale underlying the use of the survivors’ rate function or the adjusted rate function to account for the presence of the dependent terminal event. For the proposed models, we identify and develop estimation and inference procedures which can be readily implemented based on existing software. We establish the asymptotic properties of the new estimator. Simulation studies demonstrate good finite-sample performance of the proposed methods. An application to a dataset from the Cystic Fibrosis Foundation Patient Registry (CFFPR) illustrates the practical utility of the new methods.

REFERENCES


Key words and phrases. Recurrent events, counting process, Inverse probability censoring weighting, terminal event.
EARLY IDENTIFICATION OF AN IMPENDING ROCKSLIDE LOCATION VIA A SPATIALLY-AIDED GAUSSIAN MIXTURE MODEL

BY SHUO ZHOU1,*, HOWARD BONDELL1,†, ANTOINETTE TORDESILLAS1,‡, BENJAMIN I. P. RUBINSTEIN2,§ AND JAMES BAILEY2,¶

1School of Mathematics and Statistics, The University of Melbourne, *shuo.zhou@unimelb.edu.au; †howard.bondell@unimelb.edu.au; ‡atordesi@unimelb.edu.au
2School of Computing and Information Systems, The University of Melbourne, §brubinstein@unimelb.edu.au; ¶bailey@unimelb.edu.au

Movement of soil and rocks in an unstable slope under gravitational forces is an example of a complex system that is highly dynamic in space and time. A typical failure in such systems is a landslide. Fundamental studies of granular media failure combined with a complex network analysis of radar monitoring data show that distinct partitions emerge in the kinematic field in the early stages of the prefailure regime, and these patterns yield clues to the ultimate location of failure. In this study we address this partitioning of constituent units in complex systems by clustering the kinematic data, specifically, with a Gaussian mixture model. In addition, we assume that neighboring units should move together. As a result, spatial information is taken into account in our model so that spatial proximity is retained. Our case study of a rockslide from high resolution radar monitoring data shows that, by incorporating spatial information, our approach is more effective in revealing the dynamics of the system and detecting the location of a potential landslide, compared to the use of only the kinematics.

REFERENCES


Key words and phrases. Clustering, landslide, mixture model, rockslide monitoring.
A RANDOM EFFECTS STOCHASTIC BLOCK MODEL FOR JOINT COMMUNITY DETECTION IN MULTIPLE NETWORKS WITH APPLICATIONS TO NEUROIMAGING

BY SUBHADEEP PAUL\textsuperscript{1} AND YUGUO CHEN\textsuperscript{2}

\textsuperscript{1}Department of Statistics, The Ohio State University, paul.963@osu.edu
\textsuperscript{2}Department of Statistics, University of Illinois at Urbana-Champaign, yuguo@illinois.edu

To analyze data from multisubject experiments in neuroimaging studies, we develop a modeling framework for joint community detection in a group of related networks that can be considered as a sample from a population of networks. The proposed random effects stochastic block model facilitates the study of group differences and subject-specific variations in the community structure. The model proposes a putative mean community structure, which is representative of the group or the population under consideration but is not the community structure of any individual component network. Instead, the community memberships of nodes vary in each component network with a transition matrix, thus modeling the variation in community structure across a group of subjects. To estimate the quantities of interest, we propose two methods: a variational EM algorithm and a model-free “two-step” method called Co-OSNTF which is based on nonnegative matrix factorization. We also develop a resampling-based hypothesis test for differences between community structure in two populations both at the whole network level and node level. The methodology is applied to the COBRE dataset, a publicly available fMRI dataset from multisubject experiments involving schizophrenia patients. Our methods reveal an overall putative community structure representative of the group as well as subject-specific variations within each of the two groups, healthy controls and schizophrenia patients. The model has good predictive ability for predicting community structure in subjects from the same population but outside the training sample. Using our network level hypothesis tests, we are able to ascertain statistically significant difference in community structure between the two groups, while our node level tests help determine the nodes that are driving the difference.

REFERENCES


\textit{Key words and phrases.} Community detection, neuroimaging, nonnegative matrix factorization, population of networks, random effects stochastic block model.


ANALYSES OF PREVENTIVE CARE MEASURES WITH INCOMPLETE HISTORICAL DATA IN ELECTRONIC MEDICAL RECORDS: AN EXAMPLE FROM COLORECTAL CANCER SCREENING

BY YINGYE ZHENG\textsuperscript{1}, DOUGLAS A. CORLEY\textsuperscript{2}, CHYKE DOUBENI\textsuperscript{3}, ETHAN HALM\textsuperscript{4}, SUSAN M. SHORTREED\textsuperscript{5}, WILLIAM E. BARLOW\textsuperscript{6}, ANN ZAUBER\textsuperscript{7}, TOR DEVIN TOSTESON\textsuperscript{8}, AND JESSICA CHUBAK\textsuperscript{5}

\textsuperscript{1}Department of Biostatistics, Fred Hutchinson Cancer Research Center, yzheng@fredhutch.org
\textsuperscript{2}Division of Research, Kaiser Permanente Northern California, douglas.corley@kp.org
\textsuperscript{3}Department of Family Medicine, Mayo Clinic School of Medicine, Doubeni.Chyke@mayo.edu
\textsuperscript{4}Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical School, Ethan.Halm@UTSouthwestern.edu
\textsuperscript{5}Kaiser Permanente Washington Health Research Institute, Susan.M.Shortreed@kp.org
\textsuperscript{6}Cancer Research and Biostatistics, williamb@crab.org
\textsuperscript{7}Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, zaubera@MSKCC.ORG
\textsuperscript{8}Biomedical Data Science, Dartmouth College, tor.tosteson@dartmouth.edu

The calculation of quality of care measures based on electronic medical records (EMRs) may be inaccurate because of incomplete capture of past services. We evaluate the influence of different statistical approaches for calculating the proportion of patients who are up-to-date for a preventive service, using the example of colorectal cancer (CRC) screening. We propose an extension of traditional mixture models to account for the uncertainty in compliance which is further complicated by the choice of various screening modalities with different recommended screening intervals. We conducted simulation studies to compare various statistical approaches and demonstrated that the proposed method can alleviate bias when individuals with complete prior medical history information were not representative of the targeted population. The method is motivated by and applied to data from the National Cancer Institute–funded consortium Population-Based Research Optimizing Screening through Personalized Regiments (PROSPR). Findings from the application are important for the evaluation of appropriate use of preventive care and provide a novel tool for dealing with similar analytical challenges with EMR data in broad settings.

REFERENCES


Key words and phrases. Cancer screening, EMR data, event-time analysis, mixture model.
ACCOUNTING FOR DEPENDENT ERRORS IN PREDICTORS AND TIME-TO-EVENT OUTCOMES USING ELECTRONIC HEALTH RECORDS, VALIDATION SAMPLES AND MULTIPLE IMPUTATION

BY MARK J. GIGANTI1,*, PAMELA A. SHAW2, GUANHUA CHEN3, SALLY S. BEBAWY4,‡, MEGAN M. TURNER4,§, TIMOTHY R. STERLING4,¶ AND BRYAN E. SHEPHERD1,†

1 Department of Biostatistics, Vanderbilt University, *mark.giganti@vanderbilt.edu; †bryan.shepherd@vanderbilt.edu
2 Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, shawp@penmedicine.upenn.edu
3 Department of Biostatistics and Medical Informatics, University of Wisconsin, gchen25@wisc.edu
4 Vanderbilt University School of Medicine, ‡sally.furukawa@vumc.org; §megan.turner@vumc.org; ¶timothy.sterling@vanderbilt.edu

Data from electronic health records (EHR) are prone to errors which are often correlated across multiple variables. The error structure is further complicated when analysis variables are derived as functions of two or more error-prone variables. Such errors can substantially impact estimates, yet we are unaware of methods that simultaneously account for errors in covariates and time-to-event outcomes. Using EHR data from 4217 patients, the hazard ratio for an AIDS-defining event associated with a 100 cell/mm³ increase in CD4 count at ART initiation was 0.74 (95%CI: 0.68–0.80) using unvalidated data and 0.60 (95%CI: 0.53–0.68) using fully validated data. Our goal is to obtain unbiased and efficient estimates after validating a random subset of records. We propose fitting discrete failure time models to the validated subsample and then multiply imputing values for unvalidated records. We demonstrate how this approach simultaneously addresses dependent errors in predictors, time-to-event outcomes, and inclusion criteria. Using the fully validated dataset as a gold standard, we compare the mean squared error of our estimates with those from the unvalidated dataset and the corresponding subsample-only dataset for various subsample sizes. By incorporating reasonably sized validated subsamples and appropriate imputation models, our approach had improved estimation over both the naive analysis and the analysis using only the validation subsample.

REFERENCES


Key words and phrases. Correlated errors, electronic health records, measurement error, multiple imputation, survival analysis.


Causal inference from observational studies with clustered interference, with application to a cholera vaccine study ......................BRIAN G. BARKLEY, MICHAEL G. HUDGENS, JOHN D. CLEMENS, MOHAMMAD ALI AND MICHAEL E. EMCH
Quantifying time-varying sources in magnetoencephalography—a discrete approach
ZHIGANG YAO, ZENGYAN FAN, MASAHITO HAYASHI AND WILLIAM EDDY
Active matrix factorization for surveys
CHELSEA ZHANG, SEAN J. TAYLOR, CURTISS COBB AND JASJEET SEKHON
Optimal EMG placement for a robotic prosthesis controller with sequential, adaptive functional estimation ........ JONATHAN STALLRICH, MD NAZMUL ISLAM, ANA-MARIA STACIU, DUSTIN CROUCH, LIZHI PAN AND HE HUANG
Size estimation of key populations in the HIV epidemic in Eswatini using incomplete and misaligned capture-recapture data ........ ABHIRUP DATTA ANDREW PITA, AMRITA RAO, BHEKIE SITHOLE, ZANDILE MNISI AND STEFAN BARAL
A Bayesian hierarchical model for evaluating forensic footwear evidence
NEIL A. SPENCER AND JARED S. MURRAY
Statistical methods for replicability assessment ........ KENNETH HUNG AND WILLIAM FITHIAN
Statistical methods for analysis of combined categorical biomarker data from multiple studies
CHAO CHENG AND MOLIN WANG
Does terrorism trigger online hate speech? On the association of events and time series
ERIK SCHARWÄCHTER AND EMMANUEL MÜLLER
A semiparametric mixture method for local false discovery rate estimation from multiple studies
SEOK-OH JENG, DONGSEOK CHOI AND WONCHEOL JANG
Mining events with declassified diplomatic documents
YUANJUN GAO, JACK GOETZ, MATTHEW CONNELLY AND RAHUL MAZUMDER
A novel change point approach for the detection of gas emission sources using remotely contained concentration data ........ IDRIS ECKLEY, CLAUDIA KIRCH AND SILKE WEBER
PTEM: A popularity-based topical expertise model for community question answering
HOHYUN JUNG, JAE-GIL LEE, NAMGIL LEE AND SUNG-HO KIM
Spatio-temporal probabilistic wind vector forecasting over Saudi Arabia
AMANDA LENZI AND MARC G. GENTON
Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball ..................... NATHAN SANDHOLTZ AND LUKE BORRN
The Jensen effect and functional single index models: Estimating the ecological implications of nonlinear reaction norms ............ ZI YE, GILES HOOKER AND STEPHEN P. ELLNER
Efficiency in lung transplant allocation strategies
JINGJING ZOU, DAVID J. LEDERER, DANIEL RABINOWITZ
Monotonic effects of characteristics on returns
JARED DALE FISHER, DAVID WALKER PUELZ AND CARLOS MARINHO CARVALHO

Continued
The Annals of Applied Statistics

Next Issues—Continued

Hawkes binomial topic model with applications to coupled conflict-Twitter data
GEORGE MOHLER, ERIN McGRATH, CODY BUNTAIN AND GARY LAFREE

Effective model calibration via sensible variable identification and adjustment, with application to composite fuselage simulation ................. YAN WANG, XIAOWEI YUE, RUI TUO, JEFFREY HUNT AND JIANJUN SHI

A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes . . . MATTHEW KOSLOVSKY, KRISTI L. HOFFMAN, CARRIE R. DANIEL AND MARINA VANNucci
BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE
STATISTICAL
INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE