THE ANNALS of APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Articles

A Bayesian precision medicine framework for calibrating individualized therapeutic indices in cancer
ABHISEK SAHA, MIN JIN HA, SATWIK ACHARYYA AND VEERABHADRAN BALADANDAYUTHAPANI 2055

Semiparametric Bayesian forecasting of SpatioTemporal earthquake occurrences
GORDON J. ROSS AND ALEKSANDAR A. KOLEV 2083

Estimating the stillbirth rate for 195 countries using a Bayesian sparse regression model with temporal smoothing ... ZHENGFAN WANG, MIRANDA J. FIX, LUCIA HUG, ANU MISHRA, DANZHEN YOU, HANNAH BLENCOWE, JON WAKEFIELD AND LEONTINE ALKEMA 2101

Functional random effects modeling of brain shape and connectivity
EARLIL A A AND JOHN A. D. ASTON 2122

Hierarchical resampling for bagging in multistudy prediction with applications to human neurochemical sensing ........................., GABRIEL LOEWINGER, PRASAD PATIL, KENNETH T. KISHIDA AND GIOVANNI PARMIGIANI 2145

Network differential connectivity analysis .........................., SEN ZHAO AND ALI SHOAIE 2166

Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data
QIUYU WU AND XIANGYU LUO 2183

Accounting for survey design in Bayesian disaggregation of survey-based areal estimates of proportions: An application to the American Community Survey .... MARCO H. BENEDETTI, VERONICA J. BERROCAL AND RODERICK J. LITTLE 2201

Bayesian multivariate sparse functional principal components analysis with application to longitudinal microbiome multivariate data .... LINGJING JIANG, CHRIS ELMOD, JANE J. KIM, AUSTIN D. SWAFFORD, ROB KNIGHT AND WESLEY K. THOMPSON 2231

Data-adaptive efficient estimation strategies for biomarker studies embedded in randomized trials
WEI ZHANG, ZHIWEI ZHANG, JAMES F. TROENDLE AND AIYI LIU 2250

An omnibus test for detection of subgroup treatment effects via data partitioning
YIFEI SUN, XUMING HE AND JIANHUA HU 2266

Assessing treatment effect through compliance score in randomized trials with noncompliance
ZONGHUI HU, ZHIWEI ZHANG AND DEAN FOLLMANN 2279

Spatiotemporal satellite data imputation using sparse functional data analysis
WEICHENG ZHU, ZHENGYUAN ZHU AND XIONGTAO DAI 2291

Multilevel time-series models for small area estimation at different frequencies and domain levels
HARM JAN BOONSTRA AND JAN VAN DEN BRAKEL 2314

How many refugees and migrants died trying to reach Europe? Joint population size and total estimation ........................................, ALESSIO FARCOMENI 2339

Full Bayesian inference in hidden Markov models of plant growth
GAUTIER VIAUD, YUTING CHEN AND PAUL-HENRY COURNOËDE 2352

Extended stochastic block models with application to criminal networks
SIRIO LEGRAMANTI, TOMMASO RIGN, DANIELE DURANTE AND DAVID B. DUNSON 2369

continued
THE ANNALS
of
APPLIED
STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

Two-sample tests for multivariate repeated measurements of histogram objects with applications to wearable device data.......................... JINGRU ZHANG, KATHLEEN R. MERIKANGAS, HONGZHE LI AND HAOCHANG SHOU 2396

Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data
ALESSANDRO CASA, TOM F. O’CALLAGHAN AND THOMAS BRENDAN MURPHY 2417

Phylogenetically informed Bayesian truncated copula graphical models for microbial association networks ........................ HEE CHEOL CHUNG, IRINA GAYNANOVA AND YANG NI 2437

Causal inference for the effect of mobility on COVID-19 deaths
MATTEO BONVINI, EDWARD H. KENNEDY, VALERIE VENTURA AND LARRY WASSERMAN 2458

Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials
RUITAO LIN, HAOLUN SHI, GUOSHENG YIN, PETER F. THALL, YING YUAN AND CHRISTOPHER R. FLOWERS 2481

Estimating functional parameters for understanding the impact of weather and government interventions on COVID-19 outbreak....................... CHIH-LI SUNG 2505

Clustering and forecasting multiple functional time series
CHEN TANG, HAN LIN SHANG AND YANRONG YANG 2523

Semi-supervised nonparametric Bayesian modelling of spatial proteomics
OLIVER M. CROOK, KATHRYN S. LILLEY, LAURENT GATTO AND PAUL D. W. KIRK 2554

Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data
JOSEPH FELDMAN AND DANIEL R. KOWAL 2577

Hierarchical Bayesian modeling of ocean heat content and its uncertainty
SAMUEL BAUGH AND KAREN McKINNON 2603

Bayesian inference for brain activity from functional magnetic resonance imaging collected at two spatial resolutions.......................... ANDREW S. WHITEMAN, ANDREAS J. BARTSCH, JIAN KANG AND TIMOTHY D. JOHNSON 2626

Testing for differential abundance in compositional counts data, with application to microbiome studies ....................... BARAK BRILL, AMNON AMIR AND RUTH HELLER 2648

Mapping interstellar dust with Gaussian processes ........ ANDREW C. MILLER, LAUREN ANDERSON, BORIS LEISTEDT, JOHN P. CUNNINGHAM, DAVID W. HOGG AND DAVID M. Blei 2672

Modelling extremes of spatial aggregates of precipitation using conditional methods
JORDAN RICHARDS, JONATHAN A. TAWN AND SIMON BROWN 2714

A spatial causal analysis of wildland fire-contributed PM2.5 using numerical model output
ALEXANDRA LARSEN, SHU YANG, BRIAN J. REICH AND ANA G. RAPPOLD 2732

Dimensions, power and factors in an observational study of behavioral problems after physical abuse of children...................... TING YE, DYLAN S. SMALL AND PAUL R. ROSENBAUM 2755

Scalar on network regression via boosting.............. EMILY L. MORRIS, KEVIN HE AND JIAN KANG 2774

Multiscale spectral modelling for nonstationary time series within an ordered multiple-trial experiment ................ JONATHAN EMBLETON, MARINA I. KNIGHT AND HERNANDO OMBAO 2797

Bayesian bi-clustering methods with applications in computational biology
HAN YAN, JIE XING WU, YANG LI AND JUN S. LIU 2804
A BAYESIAN PRECISION MEDICINE FRAMEWORK FOR CALIBRATING INDIVIDUALIZED THERAPEUTIC INDICES IN CANCER

BY ABHISEK SAHA$^{1,a}$, MIN JIN HA$^{2,b}$, SATWIK ACHARYYA$^{3,c}$ AND VEERABHADRAN BALADANDAYUTHAPANI$^{3,d}$

1Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Health, aabhisek.saha@uconn.edu
2Department of Biostatistics, Graduate School of Public Health, Yonsei University, Seoul, Korea, bminjin.ha@gmail.com
3Department of Biostatistics, University of Michigan, csatwika@umich.edu, dveerab@umich.edu

The development and clinical implementation of evidence-based precision medicine strategies has become a realistic possibility, primarily due to the rapid accumulation of large-scale genomics and pharmacological data from diverse model systems: patients, cell lines and drug perturbation studies. We introduce a novel Bayesian modeling framework called the individualized therapeutic index (iRX) model to integrate high-throughput pharmacogenomic data across model systems. Our iRX model achieves three main goals: first, it exploits the conserved biology between patients and cell lines to calibrate therapeutic response of drugs in patients; second, it finds optimal cell line avatars as proxies for patient(s); and finally, it identifies key genomic drivers explaining cell line-patient similarities. This is achieved through a semi-supervised learning approach that conflates (unsupervised) sparse latent factor models with (supervised) penalized regression techniques. We propose a unified and tractable Bayesian model for estimation, and inference is conducted via efficient posterior sampling schemes. We illustrate and validate our approach using two existing clinical trial data sets in multiple myeloma and breast cancer studies. We show that our iRX model improves prediction accuracy compared to naive alternative approaches, and it consistently outperforms existing methods in literature in both multiple simulation scenarios as well as real clinical examples.

REFERENCES


Key words and phrases. Bayesian methods, genomic data integration, high-dimensional regression, latent factor models, precision medicine, semi-supervised learning.


SEMIPARAMETRIC BAYESIAN FORECASTING OF SPATIOTEMPORAL EARTHQUAKE OCCURRENCES

BY GORDON J. ROSS and ALEKSANDAR A. KOLEV

School of Mathematics, University of Edinburgh, \textsuperscript{a}gordon.ross@ed.ac.uk, \textsuperscript{b}aleksandar.kolev@ed.ac.uk

The Epidemic Type Aftershock Sequence (ETAS) model is a self-exciting point process which is used to model and forecast the occurrence of earthquakes in a geographical region. The ETAS model assumes that the occurrence of mainshock earthquakes follows an inhomogeneous spatial point process, with their aftershock earthquakes modelled via a separate triggering kernel. Most previous studies of the ETAS model have relied on point estimates of the model parameters, due to the complexity of the likelihood function and the difficulty in estimating an appropriate spatial mainshock distribution. In order to take estimation uncertainty into account, we instead propose a fully Bayesian formulation of the ETAS model, which uses a nonparametric Dirichlet process mixture prior to capture the spatial mainshock process, and show how efficient parameter inference can be carried out using auxiliary latent variables. We demonstrate how our model can be used for medium-term earthquake forecasts in a number of geographical regions.

REFERENCES


\begin{footnotesize}
\begin{itemize}
\item Key words and phrases. Dirichlet process, ETAS, Hawkes process, KDE, spatial analysis, Italy, Bayesian analysis, seismology, aftershocks.
\end{itemize}
\end{footnotesize}


ESTIMATING THE STILLBIRTH RATE FOR 195 COUNTRIES USING A BAYESIAN SPARSE REGRESSION MODEL WITH TEMPORAL SMOOTHING

BY ZHENGFAN WANG1,a, MIRANDA J. FIX2,c, LUCIA HUG3,e, ANU MISHRA3,f, DANZHEN YOU3,g, HANNAH BLENCOE4,h, JON WAKEFIELD2,d AND LEONTINE ALKEMA1,b

1Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, azhengfanwang@umass.edu, balkema@umass.edu
2Department of Biostatistics, University of Washington, cmiranda.fix@gmail.com, djonno@uw.edu
3Division of Data, Analytics, Planning and Monitoring, UNICEF, elhug@unicef.org, fa.mishra@imperial.ac.uk, gdyou@unicef.org
4London School of Hygiene and Tropical Medicine, University of London, hannah-jayne.blencoe@lshtm.ac.uk

Estimation of stillbirth rates globally is complicated because of the paucity of reliable data from countries where most stillbirths occur. We compiled data and developed a Bayesian hierarchical temporal sparse regression model for estimating stillbirth rates for 195 countries from 2000 to 2019. The model combines covariates with a temporal smoothing process so that estimates are data-driven in country-periods with high-quality data and determined by covariates for country-periods with limited or no data. Horseshoe priors are used to encourage sparseness. The model adjusts observations with alternative stillbirth definitions and accounts for various sources of uncertainty. In-sample goodness of fit and out-of-sample validation results suggest that the model is reasonably well calibrated. The model is used by the UN Interagency Group for Child Mortality Estimation to monitor the stillbirth rate for 195 countries.

REFERENCES


Key words and phrases. Bayesian hierarchical model, Bayesian sparsity, time-series analysis.


FUNCTIONAL RANDOM EFFECTS MODELING OF BRAIN SHAPE AND CONNECTIVITY

BY EARDI LILA1,a AND JOHN A. D. ASTON2,b

1Department of Biostatistics, University of Washington, aelila@uw.edu
2Statistical Laboratory, University of Cambridge, b.j.aston@statslab.cam.ac.uk

We present a statistical framework that jointly models brain shape and functional connectivity which are two complex aspects of the brain that have been classically studied independently. We adopt a Riemannian modeling approach to account for the non-Euclidean geometry of the space of shapes and the space of connectivity that constrains trajectories of covariation to be valid statistical estimates. In order to disentangle genetic sources of variability from those driven by unique environmental factors, we embed a functional random effects model in the Riemannian framework. We apply the proposed model to the Human Connectome Project dataset to explore spontaneous covariation between brain shape and connectivity in young healthy individuals.

REFERENCES


Key words and phrases. Functional data analysis, variance component models, mixed effects models, neuroimaging.


HIERARCHICAL RESAMPLING FOR BAGGING IN MULTISTUDY PREDICTION WITH APPLICATIONS TO HUMAN NEUROCHEMICAL SENSING

BY GABRIEL LOEWINGER\textsuperscript{1,a}, PRASAD PATIL\textsuperscript{2,b}, KENNETH T. KISHIDA\textsuperscript{3,c} AND GIOVANNI PARMIGIANI\textsuperscript{4,d}

\textsuperscript{1}Department of Biostatistics, Harvard T.H. Chan School of Public Health, \textsuperscript{a}gloewinger@gmail.com
\textsuperscript{2}Department of Biostatistics, Boston University School of Public Health, \textsuperscript{b}patil@bu.edu
\textsuperscript{3}Department of Physiology and Pharmacology, Wake Forest School of Medicine, \textsuperscript{c}kkishida@wakehealth.edu
\textsuperscript{4}Department of Data Science, Dana Farber Cancer Institute, \textsuperscript{d}gp@jimmy.harvard.edu

We propose the “study strap ensemble,” which combines advantages of two common approaches to fitting prediction models when multiple training datasets (“studies”) are available: pooling studies and fitting one model vs. averaging predictions from multiple models each fit to individual studies. The study strap ensemble fits models to bootstrapped datasets or “pseudo-studies.” These are generated by resampling from multiple studies with a hierarchical resampling scheme that generalizes the randomized cluster bootstrap. The study strap is controlled by a tuning parameter that determines the proportion of observations to draw from each study. When the parameter is set to its lowest value, each pseudo-study is resampled from only a single study. When it is high, the study strap ignores the multistudy structure and generates pseudo-studies by merging the datasets and drawing observations like a standard bootstrap. We empirically show the optimal tuning value often lies in between and prove that special cases of the study strap draw the merged dataset and the set of original studies as pseudo-studies. We extend the study strap approach with an ensemble weighting scheme that utilizes information in the distribution of the covariates of the test dataset.

Our work is motivated by neuroscience experiments using real-time neurochemical sensing during awake behavior in humans. Current techniques to perform this kind of research require measurements from an electrode placed in the brain during awake neurosurgery and rely on prediction models to estimate neurotransmitter concentrations from the electrical measurements recorded by the electrode. These models are trained by combining multiple datasets that are collected in vitro under heterogeneous conditions in order to promote accuracy of the models when applied to data collected in the brain. A prevailing challenge is deciding how to combine studies or ensemble models trained on different studies to enhance model generalizability.

Our methods produce marked improvements in simulations and in this application. All methods are available in the studyStrap CRAN package.

REFERENCES


\textit{Key words and phrases.} Domain adaptation, domain generalization, transfer learning, neuroscience.


Identifying differences in networks has become a canonical problem in many biological applications. Existing methods try to accomplish this goal by either directly comparing the estimated structures of two networks or testing the null hypothesis that the covariance or inverse covariance matrices in two populations are identical. However, estimation approaches do not provide measures of uncertainty, for example, p-values, whereas existing testing approaches could lead to misleading results, as we illustrate in this paper. To address these shortcomings, we propose a qualitative hypothesis testing framework which tests whether the connectivity structures in the two networks are the same. Our framework is especially appropriate if the goal is to identify nodes or edges that are differentially connected. No existing approach could test such hypotheses and provide corresponding measures of uncertainty. Theoretically, we show that, under appropriate conditions, our proposal correctly controls the type-I error rate in testing the qualitative hypothesis. Empirically, we demonstrate the performance of our proposal using simulation studies and applications in cancer genomics.

REFERENCES


Key words and phrases. Differential connectivity, biological networks, significance test, lasso, high-dimensional data.


ESTIMATING HETEROGENEOUS GENE REGULATORY NETWORKS FROM ZERO-INFLATED SINGLE-CELL EXPRESSION DATA

BY QIUYU WUa AND XIANGYU Luo

Institute of Statistics and Big Data, Renmin University of China, \textsuperscript{a}w.qy@ruc.edu.cn, \textsuperscript{b}xiangyulo@ruc.edu.cn

Inferring gene regulatory networks can elucidate how genes work cooperatively. The gene-gene collaboration information is often learned by Gaussian graphical models (GGM) that aim to identify whether the expression levels of any pair of genes are dependent, given other genes’ expression values. One basic assumption that guarantees the validity of GGM is data normality, and this often holds for \textit{bulk-level} expression data which aggregate biological signals from a collection of cells. However, fine-grained \textit{cell-level} expression profiles collected in single-cell RNA-sequencing (scRNA-seq) reveal non-normality features—cellular heterogeneity and zero inflation. We propose a Bayesian latent mixture GGM to jointly estimate multiple gene regulatory networks accounting for the zero inflation and unknown heterogeneity of single-cell expression data. The proposed approach outperforms competing methods on synthetic data in terms of network structure and precision matrix estimation accuracy and provides biological insights when applied to two real-world scRNA-seq datasets. An R package implementing the proposed model is available on GitHub https://github.com/WgitU/BLGGM.

REFERENCES


\textbf{Key words and phrases.} Bayesian analysis, Gaussian graphical models, heterogeneity, nonignorable dropout, spike-slab prior.
ACCOUNTING FOR SURVEY DESIGN IN BAYESIAN DISAGGREGATION OF SURVEY-BASED AREAL ESTIMATES OF PROPORTIONS: AN APPLICATION TO THE AMERICAN COMMUNITY SURVEY

BY MARCO H. BENEDETTI1,a, VERONICA J. BERROCAL2,b AND RODERICK J. LITTLE3,c

1Nationwide Children’s Hospital, Center for Injury Research and Policy, amarco.benedetti@nationwidechildrens.org
2Department of Statistics, School of Information and Computer Sciences, University of California, Irvine, bverrocal@uci.edu
3Department of Biostatistics, School of Public Health, University of Michigan, crlittle@umich.edu

Understanding the effects of social determinants of health on health outcomes requires data on characteristics of the neighborhoods in which subjects live. However, estimates of these characteristics are often aggregated over space and time in a fashion that diminishes their utility. Take, for example, estimates from the American Community Survey (ACS), a multiyear nationwide survey administered by the U.S. Census Bureau: estimates for small municipal areas are aggregated over 5-year periods, whereas 1-year estimates are only available for municipal areas with populations >65,000. Researchers may wish to use ACS estimates in studies of population health to characterize neighborhood-level exposures. However, 5-year estimates may not properly characterize temporal changes or align temporally with other data in the study, while the coarse spatial resolution of the 1-year estimates diminishes their utility in characterizing neighborhood exposure. To circumvent this issue, in this paper we propose a modeling framework to disaggregate estimates of proportions derived from sampling surveys, which explicitly accounts for the survey design effect. We illustrate the utility of our model by applying it to the ACS data, generating estimates of poverty for the state of Michigan at fine spatiotemporal resolution.

REFERENCES


Key words and phrases. Spatiotemporal change of support problem, Bayesian hierarchical model, multi-resolution approximation, latent spatiotemporal process, American Community Survey, survey-based estimates.


Microbiome researchers often need to model the temporal dynamics of multiple complex, nonlinear outcome trajectories simultaneously. This motivates our development of multivariate Sparse Functional Principal Components Analysis (mSFPCA), extending existing SFPCA methods to simultaneously characterize multiple temporal trajectories and their interrelationships. As with existing SFPCA methods, the mSFPCA algorithm characterizes each trajectory as a smooth mean plus a weighted combination of the smooth major modes of variation about the mean, where the weights are given by the component scores for each subject. Unlike existing SFPCA methods, the mSFPCA algorithm allows estimation of multiple trajectories simultaneously, such that the component scores, which are constrained to be independent within a particular outcome for identifiability, may be arbitrarily correlated with component scores for other outcomes. A Cholesky decomposition is used to estimate the component score covariance matrix efficiently and guarantee positive semidefiniteness given these constraints. Mutual information is used to assess the strength of marginal and conditional temporal associations across outcome trajectories. Importantly, we implement mSFPCA as a Bayesian algorithm using R and stan, enabling easy use of packages such as PSIS-LOO for model selection and graphical posterior predictive checks to assess the validity of mSFPCA models. Although we focus on application of mSFPCA to microbiome data in this paper, the mSFPCA model is of general utility and can be used in a wide range of real-world applications.

REFERENCES


Key words and phrases. Bayesian, functional data analysis, longitudinal, microbiome, multiomics.


DATA-ADAPTIVE EFFICIENT ESTIMATION STRATEGIES FOR BIOMARKER STUDIES EMBEDDED IN RANDOMIZED TRIALS

BY WEI ZHANG\(^1,a\), ZHIWEI ZHANG\(^2,b\), JAMES F. TROENDEL\(^3,c\) AND AIYI LIU\(^4,d\)

\(^1\)Academy of Mathematics and Systems Science, Chinese Academy of Sciences, \(^a\)zhangwei@amss.ac.cn
\(^2\)National Cancer Institute, National Institutes of Health, \(^b\)zhiwei.zhang@nih.gov
\(^3\)National Heart, Lung and Blood Institute, National Institutes of Health, \(^c\)james.troendle@nih.gov
\(^4\)Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, \(^d\)liua@mail.nih.gov

Predictive and prognostic biomarkers are increasingly important in clinical research and practice. Biomarker studies are frequently embedded in randomized clinical trials with biospecimens collected at baseline and assayed for biomarkers, either in real time or retrospectively. This article proposes efficient estimation strategies for two study settings in terms of biomarker ascertainment: a complete-data setting in which the biomarker is measured for all subjects in the trial, and a two-phase sampling design in which the biomarker is measured retrospectively for a random subsample of subjects selected in an outcome-dependent fashion. In both settings, efficient estimating functions are characterized using semiparametric theory and approximated using data-adaptive machine learning methods, leading to estimators that are consistent, asymptotically normal and (approximately) efficient under general conditions. The proposed methods are evaluated in simulation studies and applied to real data from two biomarker studies, one in each setting.

REFERENCES


**Key words and phrases.** Augmentation, precision medicine, semiparametric theory, super learner, two-phase sampling.
Late-stage clinical trials have been conducted primarily to establish the efficacy of a new treatment in an intended population. A corollary of population heterogeneity in clinical trials is that a treatment might be effective for one or more subgroups, rather than for the whole population of interest. As an example, the phase III clinical trial of panitumumab in metastatic colorectal cancer patients failed to demonstrate its efficacy in the overall population, but a subgroup associated with tumor KRAS status was found to be promising (Peeters et al. (Am. J. Clin. Oncol. 28 (2010) 4706–4713)). As we search for such subgroups via data partitioning based on a large number of biomarkers, we need to guard against inflated type I error rates due to multiple testing. Commonly-used multiplicity adjustments tend to lose power for the detection of subgroup treatment effects. We develop an effective omnibus test to detect the existence of, at least, one subgroup treatment effect, allowing a large number of possible subgroups to be considered and possibly censored outcomes. Applied to the panitumumab trial data, the proposed test would confirm a significant subgroup treatment effect. Empirical studies also show that the proposed test is applicable to a variety of outcome variables and maintains robust statistical power.

REFERENCES


Key words and phrases. Bootstrap, data partitioning, clinical trials, high-dimensional covariates, subgroup treatment effect.


ASSESSING TREATMENT EFFECT THROUGH COMPLIANCE SCORE IN RANDOMIZED TRIALS WITH NONCOMPLIANCE

BY ZONGHUI HU1,a, ZHIWEI ZHANG2,c AND DEAN FOLLMANN1,b

1National Institute of Allergy and Infectious Diseases, 2National Cancer Institute

A randomized trial is the gold standard for assessing the benefit of a treatment versus a control. When noncompliance is present, treatment effect depends on the tendency to comply—an attribute that is not directly measurable. Though the principal causal effect has been the most important for handling noncompliance, it is not immediately applicable to clinical decision-making as it targets the average effect in the latent strata of potential compliance. In this work, we propose the concept of compliance score, a linear combination of baseline characteristics, that uncovers the inherent attribute of compliance. We then assess the heterogeneous causal effect, namely, the causal effect of treatment as a function of baseline characteristics through the compliance score. A pseudo-response, along with a nonparametric estimation procedure, is proposed to ensure consistent and optimally efficient estimation. Compare to principal causal effect, the proposed effect is actionable and allows prediction of treatment effect at individual level. This work is motivated by and applied to a clinical trial to evaluate the benefit of antiretroviral regimens in HIV-infected patients.

REFERENCES


Key words and phrases. Causal inference, noncompliance, nonparametric regression, principal causal effect, principal stratification causal effect, randomized trial.


SPATIOTEMPORAL SATELLITE DATA IMPUTATION USING SPARSE FUNCTIONAL DATA ANALYSIS

BY WEICHENG ZHU\textsuperscript{a}, ZHENGYUAN ZHU\textsuperscript{b} AND XIONGTAO DAI\textsuperscript{c}

Department of Statistics, Iowa State University, \textsuperscript{a}weichzhu@amazon.com, \textsuperscript{b}zhuz@iastate.edu, \textsuperscript{c}xdai@iastate.edu

Many scientific applications and signal processing algorithms require complete satellite images. However, missing data in satellite images is very common due to various reasons such as cloud cover and sensor-specific problems. This paper introduces a general spatiotemporal satellite image imputation method based on sparse functional data analytic techniques. To handle observations consisting of a few longitudinally repeated satellite images that are themselves partially observed and noise-contaminated, we propose a multistep imputation method by following the best linear unbiased prediction principle and pooling information across all available locations and time points. Theoretical properties are established for the proposed approach under a new observation model for functional data that covers the dataset in question as a special case. Practical analysis on the Landsat data are conducted to illustrate and validate our algorithm which also shows that the proposed method considerably outperforms existing algorithms in terms of prediction accuracy. An efficient implementation using R and Rcpp is made available in the R package \texttt{stfit}.

REFERENCES


MICROSOFT and WESTON, S. (2020). \texttt{Foreach}: Provides \texttt{Foreach Looping Construct}. R package version 1.5.0.

\textit{Key words and phrases.} Satellite, imputation, sparse functional data, spatiotemporal, gap-fill, STFIT, R.
MULTILEVEL TIME-SERIES MODELS FOR SMALL AREA ESTIMATION AT DIFFERENT FREQUENCIES AND DOMAIN LEVELS

BY HARM JAN BOONSTRA¹,a AND JAN VAN DEN BRAKEL²,b

¹Statistics Netherlands, ²Maastricht University.

A small area estimation method is developed for repeatedly conducted multipurpose surveys. A multilevel time-series model is proposed that uses direct estimates for the most detailed domains observed at the highest frequency of the repeated survey. A consistent set of estimates at different aggregation levels is then derived by aggregation of the model-based predictions obtained for the most detailed domains observed at the highest frequency. The model borrows strength over time and space via smooth and local level trends at different aggregation levels. The model also borrows information from auxiliary series available from registers with coefficients that can vary over both domains and time. Regional domain random effects are allowed to vary smoothly over space according to a spatial autoregressive process. To account for the diversity of domains and for more volatile time-dependence, nonnormally distributed random effects and trend innovations are used via so-called global-local shrinkage priors. A Bayesian approach is taken, and the model is estimated by MCMC simulation. The method is illustrated with an application to the Dutch Labour Force Survey to produce monthly provincial and quarterly municipal unemployment figures.

REFERENCES


Key words and phrases. Hierarchical Bayesian model, global-local shrinkage, Gibbs sampler, labour force survey.


HOW MANY REFUGEES AND MIGRANTS DIED TRYING TO REACH EUROPE? JOINT POPULATION SIZE AND TOTAL ESTIMATION

BY ALESSIO FARCOMENI

Department of Economics and Finance, University of Rome “Tor Vergata”, alessio.farcomeni@uniroma2.it

We estimate the number of migrants and refugees that died while trying to enter the European Union, during a period of 25 years. Only a subset of attempts with at least one casualty are reported by at least one media source. In order to obtain the estimate, we propose a regression-extrapolation approach, for joint estimation of population size (here, the number of deadly individual or group attempts) and the sum of an accompanying trait (here, the number of deaths) over the population. The trait is measured only for a biased sample of individuals, that are repeatedly observed. Closed-form expressions are derived for the estimator and its standard error. Our findings are that about 40,000 have died from January 1993 to March 2019, during about 5500 attempts to enter the European Union. The number of deaths has been steadily increasing over time, and so has the number of deaths per attempt. About 20% of attempts with at least one casualty have not been recorded by any media source, and slightly less than 10% of deaths have thus been overlooked by media.

REFERENCES


Key words and phrases. Capture–recapture, Chao estimator, Horvitz–Thompson estimator, migration, total estimator.


FULL BAYESIAN INFERENCE IN HIDDEN MARKOV MODELS OF PLANT GROWTH

BY GAUTIER VIAUD¹,a, YUTING CHEN²,c AND PAUL-HENRY COUNNÈDE¹,b

¹Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systèmes,  
agautier.viaud@centralesupelec.fr, bpaul-henry.cournede@centralesupelec.fr  
2Energy Analysis and Environmental Impacts Division, Lawrence Berkeley National Laboratory, cyutingchen@lbl.gov

Accurately modeling the growth process of plants in interaction with their environment is important for predicting their biophysical characteristics, referred to as phenotype prediction. Most models are described by discrete dynamic systems in general state-space representation with important domain-specific characteristics: First, plant model parameters have usually clear functional meanings and may be of genetic origins, thus necessitating a precise estimation. Second, critical growth variables, specifically biomass production and dynamic allocation to organs, are hidden variables not accessible to measure. Finally, the difficulty to assess the local plant environment may imply the introduction of process noises in models. Therefore, a precise understanding of the system’s behavior requires the joint estimation of functional parameters, hidden states, and noise parameters. In this paper we describe how a full Bayesian method of estimation can accurately estimate all these key model variables using Markov chain Monte Carlo (MCMC) techniques. In the presence of both process and observation noises, it requires to use adequate particle MCMC (PMCMC) algorithms to efficiently sample the hidden states which, consequently, allows for a precise estimation of all noise parameters involved. Thanks to the Bayesian framework, appropriate choices of prior distributions for the noise parameters have enabled analytical posterior distributions and only simple updates are required. Furthermore, this estimation strategy can be easily generalized and adapted to different types of plant growth models, such as organ-scale or compartmental, provided that they are formulated as hidden Markov models. Our estimation method improves on those classically used in plant growth modeling in several aspects: First, by building upon a general probabilistic framework the estimation results allow proper statistical analyses. It is useful in prediction, not only for uncertainty and risk analysis (e.g., for crop yield prediction) but also to analyze the results of experimental trials, for example, to compare genotypes in breeding. Moreover, the care taken in the estimation of hidden variables opens new perspectives in the understanding of inner growth processes, notably the balance and interaction between biomass production and allocation (referred to as source-sink dynamics). Applications of this estimation procedure are demonstrated on the GreenLab model for Arabidopsis thaliana and the Log-Normal Allocation and Senescence (LNAS) model for sugar beet, on both synthetic and real data.

REFERENCES


Key words and phrases. Bayesian inference, hidden Markov model, plant growth, particle MCMC.
EXTENDED STOCHASTIC BLOCK MODELS WITH APPLICATION TO CRIMINAL NETWORKS

BY SIRIO LEGRAMANTI¹,a, TOMMASO RIGON²,c, DANIELE DURANTE¹,b AND DAVID B. DUNSON³,d

¹Department Decision Sciences and Institute for Data Science and Analytics, Bocconi University, 
auntsie.legramanti@unibocconi.it, b daniele.durante@unibocconi.it
²Department of Economics, Management and Statistics, University of Milano-Bicocca, c tommaso.rigon@unimib.it
³Department of Statistical Science, Duke University, d dunson@duke.edu

Reliably learning group structures among nodes in network data is challenging in several applications. We are particularly motivated by studying covert networks that encode relationships among criminals. These data are subject to measurement errors, and exhibit a complex combination of an unknown number of core-periphery, assortative and disassortative structures that may unveil key architectures of the criminal organization. The coexistence of these noisy block patterns limits the reliability of routinely-used community detection algorithms, and requires extensions of model-based solutions to realistically characterize the node partition process, incorporate information from node attributes, and provide improved strategies for estimation and uncertainty quantification. To cover these gaps, we develop a new class of extended stochastic block models (ESBM) that infer groups of nodes having common connectivity patterns via Gibbs-type priors on the partition process. This choice encompasses many realistic priors for criminal networks, covering solutions with fixed, random and infinite number of possible groups, and facilitates the inclusion of node attributes in a principled manner. Among the new alternatives in our class, we focus on the Gnedin process as a realistic prior that allows the number of groups to be finite, random and subject to a reinforcement process coherent with criminal networks. A collapsed Gibbs sampler is proposed for the whole ESBM class, and refined strategies for estimation, prediction, uncertainty quantification and model selection are outlined. The ESBM performance is illustrated in realistic simulations and in an application to an Italian mafia network, where we unveil key complex block structures, mostly hidden from state-of-the-art alternatives.

REFERENCES


Key words and phrases. Bayesian nonparametrics, Gibbs-type prior, network, product partition model.


TWO-SAMPLE TESTS FOR MULTIVARIATE REPEATED MEASUREMENTS OF HISTOGRAM OBJECTS WITH APPLICATIONS TO WEARABLE DEVICE DATA

BY JINGRU ZHANG1,a, KATHLEEN R. MERIKANGAS2,d, HONGZHE LI1,b AND HAOCHEANG SHOU1,c

1Division of Biostatistics, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, ajingru.zhang@pennmedicine.upenn.edu, bhongzhe@pennmedicine.upenn.edu

2Genetic Epidemiology Research Branch, National Institute of Mental Health, National Institutes of Health, dmerikank@mail.nih.gov

Repeated observations have become increasingly common in biomedical research and longitudinal studies. For instance, wearable sensor devices are deployed to continuously track physiological and biological signals from each individual over multiple days. It remains of great interest to appropriately evaluate how the daily distribution of biosignals might differ across disease groups and demographics. Hence, these data could be formulated as multivariate complex object data, such as probability densities, histograms, and observations on a tree. Traditional statistical methods would often fail to apply, as they are sampled from an arbitrary non-Euclidean metric space. In this paper we propose novel, nonparametric, graph-based two-sample tests for object data with the same structure of repeated measures. We treat the repeatedly measured object data as multivariate object data, which requires the same number of repeated observations per individual but eliminates any assumptions on the errors of the repeated observations. A set of test statistics are proposed to capture various possible alternatives. We derive their asymptotic null distributions under the permutation null. These tests exhibit substantial power improvements over the existing methods while controlling the type I errors under finite samples as shown through simulation studies. The proposed tests are demonstrated to provide additional insights on the location, inter- and intra-individual variability of the daily physical activity distributions in a sample of studies for mood disorders.

REFERENCES


Key words and phrases. Graph-based test, nonparametric test, non-Euclidean data, repeated measures, wearable device data.


PARSIMONIOUS BAYESIAN FACTOR ANALYSIS FOR MODELLING LATENT STRUCTURES IN SPECTROSCOPY DATA

BY ALESSANDRO CASA1,3,a, TOM F. O’CALLAGHAN2,c AND THOMAS BRENDAN MURPHY1,3,b

1School of Mathematics & Statistics, University College Dublin, alessandro.casa@ucd.ie,
brendan.murphy@ucd.ie
2School of Food & Nutritional Sciences, University College Cork, tom_ocallaghan@ucc.ie
3Vistamilk SFI Research Centre

In recent years, within the dairy sector, animal diet and management practices have been receiving increased attention, in particular, examining the impact of pasture-based feeding strategies on the composition and quality of milk and dairy products in line with the prevalence of premium grass-fed dairy products appearing on market shelves. To date, methods to thoroughly investigate the more relevant differences induced by the diet on milk chemical features are limited; enhanced statistical tools exploring these differences are required.

Infrared spectroscopy techniques are widely used to collect data on milk samples and to predict milk related traits and characteristics. While these data are routinely used to predict the composition of the macro components of milk, each spectrum also provides a reservoir of unharnessed information about the sample. The accumulation and subsequent interpretation of these data present some challenges due to their high-dimensionality and the relationships amongst the spectral variables.

In this work, directly motivated by a dairy application, we propose a modification of the standard factor analysis to induce a parsimonious summary of spectroscopic data. Our proposal maps the observations into a low-dimensional latent space while simultaneously clustering the observed variables. The method indicates possible redundancies in the data, and it helps disentangle the complex relationships among the wavelengths. A flexible Bayesian estimation procedure is proposed for model fitting, providing reasonable values for the number of latent factors and clusters. The method is applied on milk mid-infrared (MIR) spectroscopy data from dairy cows on distinctly different pasture and nonpasture based diets, providing accurate modelling of the correlation, clustering of variables, and information on differences among milk samples from cows on different diets.

REFERENCES


Key words and phrases. Dairy science, spectroscopy, chemometrics, factor analysis, redundant variables, clustering, Gibbs sampling.


Casa, A., O’Callaghan, T. F. and Murphy, T. B. (2022). Supplement to “Parsimonious Bayesian factor analysis for modelling latent structures in spectroscopy data.” [https://doi.org/10.1214/21-AOAS1597SUPP](https://doi.org/10.1214/21-AOAS1597SUPP)


PHYLOGENETICALLY INFORMED BAYESIAN TRUNCATED COPULA GRAPHICAL MODELS FOR MICROBIAL ASSOCIATION NETWORKS

BY HEE CHEOL CHUNG\textsuperscript{a}, IRINA GAYNANOVA\textsuperscript{b} AND YANG NI\textsuperscript{c}

Department of Statistics, Texas A\&M University, \textsuperscript{a}hcchung@stat.tamu.edu, \textsuperscript{b}irinag@stat.tamu.edu, \textsuperscript{c}yni@stat.tamu.edu

Microorganisms play critical roles in host health. The advancement of high-throughput sequencing technology provides opportunities for a deeper understanding of microbial interactions. However, due to the technological limitations of 16S ribosomal RNA sequencing, microbiome data are zero-inflated, and a quantitative comparison of microbial abundances cannot be made across subjects. By leveraging a recent microbiome profiling technique that quantifies 16S ribosomal RNA microbial counts, we propose a novel Bayesian graphical model that incorporates microorganisms’ evolutionary history through a phylogenetic tree prior and explicitly accounts for zero inflation using the truncated Gaussian copula. Our simulation study reveals that the evolutionary information substantially improves the network estimation accuracy. We apply the proposed model to the quantitative gut microbiome data of 106 healthy subjects and identify three distinct microbial communities that are not found by existing microbial network estimation models. We further find that these communities are discriminated based on microorganisms’ ability to utilize oxygen as an energy source.

REFERENCES


Key words and phrases. Gaussian copula, Markov random field, phylogenetic tree, zero inflation.
CAUSAL INFERENCE FOR THE EFFECT OF MOBILITY ON COVID-19 DEATHS

BY MATTEO BONVINI, EDWARD H. KENNEDY, VALERIE VENTURA AND LARRY WASSERMAN

Department of Statistics and Data Science, Carnegie Mellon University, a mbonvini@stat.cmu.edu, b edward@stat.cmu.edu, c vventura@stat.cmu.edu, d larry@stat.cmu.edu

In this paper we develop statistical methods for causal inference in epidemics. Our focus is in estimating the effect of social mobility on deaths in the first year of the Covid-19 pandemic. We propose a marginal structural model motivated by a basic epidemic model. We estimate the counterfactual time series of deaths under interventions on mobility. We conduct several types of sensitivity analyses. We find that the data support the idea that reduced mobility causes reduced deaths, but the conclusion comes with caveats. There is evidence of sensitivity to model misspecification and unmeasured confounding which implies that the size of the causal effect needs to be interpreted with caution. While there is little doubt the effect is real, our work highlights the challenges in drawing causal inferences from pandemic data.

REFERENCES


BAYESIAN HIERARCHICAL RANDOM-EFFECTS META-ANALYSIS AND DESIGN OF PHASE I CLINICAL TRIALS

BY RUITAO LIN1,a, HAOLUN SHI2,d, GUOSHENG YIN3,e, PETER F. THALL1,b, YING YUAN1,c AND CHRISTOPHER R. FLOWERS4,f

1Department of Biostatistics, The University of Texas MD Anderson Cancer Center, arlin@mdanderson.org, brex@mdanderson.org, cyyuan@mdanderson.org
2Department of Statistics and Actuarial Science, Simon Fraser University, dhaoluns@sfu.ca
3Department of Statistics and Actuarial Science, The University of Hong Kong, egyin@hku.hk
4Department of Lymphoma/Myeloma, The University of Texas MD Anderson Cancer Center, crflowers@mdanderson.org

We propose a curve-free random-effects meta-analysis approach to combining data from multiple phase I clinical trials to identify an optimal dose. Our method accounts for between-study heterogeneity that may stem from different study designs, patient populations, or tumor types. We also develop a meta-analytic-predictive (MAP) method, based on a power prior, that incorporates data from multiple historical studies into the design and conduct of a new phase I trial. Performances of the proposed methods for data analysis and trial design are evaluated by extensive simulation studies. The proposed random-effects meta-analysis method provides more reliable dose selection than comparators that rely on parametric assumptions. The MAP-based dose-finding designs are generally more efficient than those that do not borrow information, especially when the current and historical studies are similar. The proposed methodologies are illustrated by a meta-analysis of five historical phase I studies of Sorafenib and design of a new phase I trial.

REFERENCES


Key words and phrases. Bayesian adaptive method, meta-analysis, phase I clinical trials, power prior, random-effects model.


ESTIMATING FUNCTIONAL PARAMETERS FOR UNDERSTANDING THE IMPACT OF WEATHER AND GOVERNMENT INTERVENTIONS ON COVID-19 OUTBREAK

By Chih-Li Sung

Department of Statistics and Probability, Michigan State University, sungchih@msu.edu

As the coronavirus disease 2019 (COVID-19) has shown profound effects on public health and the economy worldwide, it becomes crucial to assess the impact on the virus transmission and develop effective strategies to address the challenge. A new statistical model, derived from the SIR epidemic model with functional parameters, is proposed to understand the impact of weather and government interventions on the virus spread in the presence of asymptomatic infections among eight metropolitan areas in the United States. The model uses Bayesian inference with Gaussian process priors to study the functional parameters nonparametrically, and sensitivity analysis is adopted to investigate the main and interaction effects of these factors. This analysis reveals several important results, including the potential interaction effects between weather and government interventions, which shed new light on the effective strategies for policymakers to mitigate the COVID-19 outbreak.

REFERENCES


Key words and phrases. Basic reproduction number, asymptomatic infections, epidemic model, nonparametric regression, sensitivity analysis.


CLUSTERING AND FORECASTING MULTIPLE FUNCTIONAL TIME SERIES

BY CHEN TANG\textsuperscript{1,a}, HAN LIN SHANG\textsuperscript{2,c} AND YANRONG YANG\textsuperscript{1,b}

\textsuperscript{1}Research school of Finance, Actuarial Studies and Statistics, The Australian National University, \textsuperscript{2}Department of Actuarial Studies and Business Analytics, Macquarie University, \textsuperscript{a}chen.tang@anu.edu.au, \textsuperscript{b}yanrong.yang@anu.edu.au

Modeling and forecasting homogeneous age-specific mortality rates of multiple countries could lead to improvements in long-term forecasting. Data fed into joint models are often grouped according to nominal attributes, such as geographic regions, ethnic groups, and socioeconomic status, which may still contain heterogeneity and deteriorate the forecast results. Our paper proposes a novel clustering technique to pursue homogeneity among multiple functional time series, based on functional panel data modeling, to address this issue. Using a functional panel data model with fixed effects, we can extract common functional time series features. These common features could be decomposed into two components: the functional time trend and the mode of variations of functions (functional pattern). The functional time trend reflects the dynamics across time, while the functional pattern captures the fluctuations within curves. The proposed clustering method searches for homogeneous age-specific mortality rates of multiple countries by accounting for both the modes of variations and the temporal dynamics among curves. We demonstrate that the proposed clustering technique outperforms other existing methods through a Monte Carlo simulation and could handle complicated cases with slow decaying eigenvalues. In empirical data analysis we find that the clustering results of age-specific mortality rates can be explained by the combination of geographic region, ethnic groups, and socioeconomic status. We further show that our model produces more accurate forecasts than several benchmark methods in forecasting age-specific mortality rates.

REFERENCES


Key words and phrases. Functional panel data, multilevel functional data, functional time series, functional principal component analysis, age-specific mortality forecasting.


SEMI-SUPERVISED NONPARAMETRIC BAYESIAN MODELLING OF SPATIAL PROTEOMICS

BY OLIVER M. CROOK\textsuperscript{1,2,}\textsuperscript{a}, KATHRYN S. LILLEY\textsuperscript{2,}\textsuperscript{c}, LAURENT GATTO\textsuperscript{3,}\textsuperscript{d} AND PAUL D. W. KIRK\textsuperscript{1,}\textsuperscript{b}

\textsuperscript{1}MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, \textsuperscript{a}oliver.crook@stats.ox.ac.uk, \textsuperscript{b}paul.kirk@mrc-bsu.cam.ac.uk
\textsuperscript{2}Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, \textsuperscript{c}k.s.lilley@bioc.cam.ac.uk
\textsuperscript{3}de Duve Institute, UCLouvain, \textsuperscript{d}laurent.gatto@uclouvain.be

Understanding subcellular protein localisation is an essential component in the analysis of context specific protein function. Recent advances in quantitative mass-spectrometry (MS) have led to high-resolution mapping of thousands of proteins to subcellular locations within the cell. Novel modelling considerations to capture the complex nature of these data are thus necessary. We approach analysis of spatial proteomics data in a nonparametric Bayesian framework, using \textit{K}-component mixtures of Gaussian process regression models. The Gaussian process regression model accounts for correlation structure within a subcellular niche, with each mixture component capturing the distinct correlation structure observed within each niche. The availability of \textit{marker proteins} (i.e., proteins with a priori known labelled locations) motivates a semi-supervised learning approach to inform the Gaussian process hyperparameters. We moreover provide an efficient Hamiltonian-within-Gibbs sampler for our model. Furthermore, we reduce the computational burden associated with inversion of covariance matrices by exploiting the structure in the covariance matrix. A tensor decomposition of our covariance matrices allows extended Trench and Durbin algorithms to be applied to reduce the computational complexity of inversion and hence accelerate computation. We provide detailed case-studies on \textit{Drosophila} embryos and mouse pluripotent embryonic stem cells to illustrate the benefit of semi-supervised functional Bayesian modelling of the data.

REFERENCES


\textit{Key words and phrases.} Proteomics, Bayesian mixture models, semi-supervised learning.


Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data

By Joseph Feldman\textsuperscript{a} and Daniel R. Kowal\textsuperscript{b}

Department of Statistics, Rice University, \textsuperscript{a}joseph.r.feldman@rice.edu, \textsuperscript{b}daniel.kowal@rice.edu

Much of the microdata used for epidemiological studies contain sensitive measurements on real individuals. As a result, such microdata cannot be published out of privacy concerns, and without public access to these data, any statistical analyses originally published on them are nearly impossible to reproduce. To promote the dissemination of key datasets for analysis without jeopardizing the privacy of individuals, we introduce a cohesive Bayesian framework for the generation of fully synthetic high-dimensional microdatasets of mixed categorical, binary, count, and continuous variables. This process centers around a joint Bayesian model that is simultaneously compatible with all of these data types, enabling the creation of mixed synthetic datasets through posterior predictive sampling. Furthermore, a focal point of epidemiological data analysis is the study of conditional relationships between various exposures and key outcome variables through regression analysis. We design a modified data synthesis strategy to target and preserve these conditional relationships, including both nonlinearities and interactions. The proposed techniques are deployed to create a synthetic version of a confidential dataset containing dozens of health, cognitive, and social measurements on nearly 20,000 North Carolina children.

References


Key words and phrases. Copula, data privacy, factor model, nonparametric regression.
HIERARCHICAL BAYESIAN MODELING OF OCEAN HEAT CONTENT
AND ITS UNCERTAINTY

BY SAMUEL BAUGH1,a AND KAREN MCKINNON2,b

1Department of Statistics, UCLA, a samuelbaugh@ucla.edu
2Department of Statistics, UCLA and Institute for the Environment and Sustainability, UCLA, b kmckinnon@ucla.edu

The accurate quantification of changes in the heat content of the world’s oceans is crucial for our understanding of the effects of increasing greenhouse gas concentrations. The Argo program, consisting of Lagrangian floats that measure vertical temperature profiles throughout the global ocean, has provided a wealth of data from which to estimate ocean heat content. However, creating a globally consistent statistical model for ocean heat content remains challenging due to the need for a globally valid covariance model that can capture complex nonstationarity. In this paper, we develop a hierarchical Bayesian Gaussian process model that uses kernel convolutions with cylindrical distances to allow for spatial nonstationarity in all model parameters while using a Vecchia process to remain computationally feasible for large spatial datasets. Our approach can produce valid credible intervals for globally integrated quantities that would not be possible using previous approaches. These advantages are demonstrated through the application of the model to Argo data, yielding credible intervals for the spatially varying trend in ocean heat content that accounts for both the uncertainty induced from interpolation and from estimating the mean field and other parameters. Through cross-validation, we show that our model outperforms an out-of-the-box approach as well as other simpler models. The code for performing this analysis is provided as the R package BayesianOHC.

REFERENCES


Key words and phrases. Hierarchical Bayesian modeling, ocean heat content, nonstationary spatial modeling.


BAYESIAN INFERENCE FOR BRAIN ACTIVITY FROM FUNCTIONAL MAGNETIC RESONANCE IMAGING COLLECTED AT TWO SPATIAL RESOLUTIONS

BY ANDREW S. WHITEMAN¹,a, ANDREAS J. BARTSCH²,d, JIANG KANG¹,b AND TIMOTHY D. JOHNSON¹,c

¹Department of Biostatistics, University of Michigan School of Public Health, awhitem@umich.edu, bjiankang@umich.edu, ctdjtdj@umich.edu
²Radiologie Bamberg and Department of Neuroradiology, University of Heidelberg, dbartsch@radiologie-bamberg.de

Neuroradiologists and neurosurgeons increasingly opt to use functional magnetic resonance imaging (fMRI) to map functionally relevant brain regions for noninvasive presurgical planning and intraoperative neuronavigation. This application requires a high degree of spatial accuracy, but the fMRI signal-to-noise ratio (SNR) decreases as spatial resolution increases. In practice, fMRI scans can be collected at multiple spatial resolutions, and it is of interest to make more accurate inference on brain activity by combining data with different resolutions. To this end, we develop a new Bayesian model to leverage both better anatomical precision in high resolution fMRI and higher SNR in standard resolution fMRI. We assign a Gaussian process prior to the mean intensity function and develop an efficient, scalable posterior computation algorithm to integrate both sources of data. We draw posterior samples using an algorithm analogous to Riemann manifold Hamiltonian Monte Carlo in an expanded parameter space. We illustrate our method in analysis of presurgical fMRI data and show in simulation that it infers the mean intensity more accurately than alternatives that use either the high or standard resolution fMRI data alone.

REFERENCES


Key words and phrases. Imaging statistics, Gaussian process, Bayesian nonparametrics, data integration, presurgical fMRI.


TESTING FOR DIFFERENTIAL ABUNDANCE IN COMPOSITIONAL COUNTS DATA, WITH APPLICATION TO MICROBIOME STUDIES

BY BARAK BRILL\textsuperscript{1, a}, AMNON AMIR\textsuperscript{2, c} AND RUTH HELLER\textsuperscript{1, b}

\textsuperscript{1}Department of Statistics and Operations Research, Tel Aviv University, \textsuperscript{a}barakbri@mail.tau.ac.il, \textsuperscript{b}ruheller@gmail.com
\textsuperscript{2}Microbiome center, The Chaim Sheba Medical Center, \textsuperscript{c}amnonim@gmail.com

Identifying which taxa in our microbiota are associated with traits of interest is important for advancing science and health. However, the identification is challenging because the measured vector of taxa counts (by amplicon sequencing) is compositional, so a change in the abundance of one taxon in the microbiota induces a change in the number of sequenced counts across all taxa. The data are typically sparse, with many zero counts present either due to biological variance or limited sequencing depth. We examine the case of Crohn’s disease, where the microbial load changes substantially with the disease. For this representative example of a highly compositional setting, we show existing methods designed to identify differentially abundant taxa may have an inflated number of false positives. We introduce a novel nonparametric approach that provides valid inference, even when the fraction of zero counts is substantial. Our approach uses a set of reference taxa that are non-differentially abundant which can be estimated from the data or from outside information. Our approach also allows for a novel type of testing: multivariate tests of differential abundance over a focused subset of the taxa. Genera-level multivariate testing discovers additional genera as differentially abundant by avoiding agglomeration of taxa.

REFERENCES


Key words and phrases. Compositional bias, analysis of composition, normalization, rarefaction, nonparametric tests.


MAPPING INTERSTELLAR DUST WITH GAUSSIAN PROCESSES

BY ANDREW C. MILLER¹,a, LAUREN ANDERSON⁴,e, BORIS LEISTEDT⁴,f, JOHN P. CUNNINGHAM²,c, DAVID W. HOGG³,d AND DAVID M. BLEI¹,b

¹Data Science Institute, Columbia University, ²Department of Statistics, Columbia University, ³Department of Physics, New York University,
⁴Center for Computational Astrophysics, Flatiron Institute

Interstellar dust corrupts nearly every stellar observation and accounting for it is crucial to measuring physical properties of stars. We model the dust distribution as a spatially varying latent field with a Gaussian process (GP) and develop a likelihood model and inference method that scales to millions of astronomical observations. Modeling interstellar dust is complicated by two factors. The first is integrated observations. The data come from a vantage point on Earth, and each observation is an integral of the unobserved function along our line of sight, resulting in a complex likelihood and a more difficult inference problem than in classical GP inference. The second complication is scale; stellar catalogs have millions of observations. To address these challenges, we develop ZIGGY, a scalable approach to GP inference with integrated observations based on stochastic variational inference. We study ZIGGY on synthetic data and the Ananke dataset, a high-fidelity mechanistic model of the Milky Way with millions of stars. ZIGGY reliably infers the spatial dust map with well-calibrated posterior uncertainties.

REFERENCES


Key words and phrases. Gaussian process, interstellar dust, astrostatistics, stochastic variational inference.


MODELLING EXTREMES OF SPATIAL AGGREGATES OF PRECIPITATION USING CONDITIONAL METHODS

BY JORDAN RICHARDS¹,a, JONATHAN A. TAWN¹,b AND SIMON BROWN²,c

¹STOR-i Centre for Doctoral Training, Department of Mathematics and Statistics, Lancaster University, a jordan.richards@kaust.edu.sa, b j.tawn@lancaster.ac.uk
²Hadley Centre for Climate Prediction and Research, Met Office, c simon.brown@metoffice.gov.uk

Inference on the extremal behaviour of spatial aggregates of precipitation is important for quantifying river flood risk. There are two classes of previous approach, with one failing to ensure self-consistency in inference across different regions of aggregation and the other imposing highly restrictive assumptions. To overcome these issues, we propose a model for high-resolution precipitation data from which we can simulate realistic fields and explore the behaviour of spatial aggregates. Recent developments have seen spatial extensions of the Heffernan and Tawn (J. R. Stat. Soc. Ser. B. Stat. Methodol. 66 (2004) 497–546) model for conditional multivariate extremes which can handle a wide range of dependence structures. Our contribution is twofold: extensions and improvements of this approach and its model inference for high-dimensional data and a novel framework for deriving aggregates addressing edge effects and subregions without rain. We apply our modelling approach to gridded East Anglia, UK precipitation data. Return-level curves for spatial aggregates over different regions of various sizes are estimated and shown to fit very well to the data.

REFERENCES


Key words and phrases. Extremal dependence, extreme precipitation, spatial aggregates, spatial conditional extremes.


A SPATIAL CAUSAL ANALYSIS OF WILDLAND FIRE-CONTRIBUTED PM$_{2.5}$ USING NUMERICAL MODEL OUTPUT

BY ALEXANDRA LARSEN$^{1,a}$, SHU YANG$^{2,b}$, BRIAN J. REICH$^{2,c}$ AND ANA G. RAPPOLD$^{3,d}$

$^1$Department of Biostatistics and Bioinformatics, Duke University, $^a$alexandra.larsen@duke.edu
$^2$Department of Statistics, North Carolina State University, $^b$syang24@ncsu.edu, $^c$bjreich@ncsu.edu
$^3$National Health and Environmental Effects Research Laboratory—Environmental Public Health Division, US Environmental Protection Agency, $^d$Rappold.Ana@epa.gov

Wildland fire smoke contains hazardous levels of fine particulate matter (PM$_{2.5}$), a pollutant shown to adversely affect health. Estimating fire-attributable PM$_{2.5}$ concentrations is key to quantifying the impact on air quality and subsequent health burden. This is a challenging problem since only total PM$_{2.5}$ is measured at monitoring stations and both fire-attributable PM$_{2.5}$ and PM$_{2.5}$ from all other sources are correlated in space and time. We propose a framework for estimating fire-contributed PM$_{2.5}$ and PM$_{2.5}$ from all other sources using a novel causal inference framework and bias-adjusted chemical model representations of PM$_{2.5}$ under counterfactual scenarios. The chemical model representation of PM$_{2.5}$ for this analysis is simulated using Community Multiscale Air Quality Modeling System (CMAQ), run with and without fire emissions across the contiguous U.S. for the 2008–2012 wildfire seasons. The CMAQ output is calibrated with observations from monitoring sites for the same spatial domain and time period. We use a Bayesian model that accounts for spatial variation to estimate the effect of wildland fires on PM$_{2.5}$ and state assumptions under which the estimate has a valid causal interpretation. Our results include estimates of the contributions of wildfire smoke to PM$_{2.5}$ for the contiguous U.S. Additionally, we compute the health burden associated with the PM$_{2.5}$ attributable to wildfire smoke.

REFERENCES


Key words and phrases. Bayesian analysis, downscaling, interference, spillover effect.
Many observational studies assess the impact of a treatment on an outcome that has several dimensions. In the observational study that we discuss, physical abuse of children may affect the degree to which the child exhibits depression, withdrawal or aggression. A treatment may affect all, some or none of these dimensions. In addition to the scientific interest in learning the effect on each dimension, it is also known that an appropriate combination of dimensions may increase power, efficiency and insensitivity to unmeasured biases; however, finding this appropriate combination requires corrections for multiple testing that erode power. We explore this trade-off by developing a new formula for the power of a sensitivity analysis in a simple situation with several dimensions. The methodology is applied to study the effects of physical abuse in early childhood and its possible effects on several dimensions of subsequent behavioral problems. Also, a general method is proposed for converting any signed rank test for matched pairs into an analogous test for matching each treated individual to several controls, and the performance of this extension is examined. The proposed method aids in studying the relative magnitude of the effect on different dimensions. A second evidence factor considers the dose or intensity of physical abuse.

REFERENCES


Key words and phrases. Causal inference, coherence among multiple outcomes, design sensitivity, evidence factors, observational study, power of a sensitivity analysis, Scheffé correction, sensitivity analysis.


SCALAR ON NETWORK REGRESSION VIA BOOSTING

BY EMILY L. MORRIS\textsuperscript{a}, KEVIN HE\textsuperscript{b} and JIAN KANG\textsuperscript{c}

Department of Biostatistics, University of Michigan, \textsuperscript{a}emorrisl@umich.edu, \textsuperscript{b}kevinhe@umich.edu, \textsuperscript{c}jiankang@umich.edu

Neuroimaging studies have a growing interest in learning the association between the individual brain connectivity networks and their clinical characteristics. It is also of great interest to identify the sub-brain networks as biomarkers to predict the clinical symptoms, such as disease status, potentially providing insight on neuropathology. This motivates the need for developing a new type of regression model where the response variable is scalar, and predictors are networks that are typically represented as adjacent matrices or weighted adjacent matrices to which we refer as scalar-on-network regression. In this work we develop a new boosting method for model fitting with subnetwork markers selection. Our approach, as opposed to group lasso or other existing regularization methods, is, essentially, a gradient descent algorithm leveraging known network structure. We demonstrate the utility of our methods via simulation studies and analysis of the resting-state fMRI data in a cognitive developmental cohort study.

REFERENCES


\textbf{Key words and phrases.} Neuroimaging, fMRI, boosting.


MULTISCALE SPECTRAL MODELLING FOR NONSTATIONARY TIME SERIES WITHIN AN ORDERED MULTIPLE-TRIAL EXPERIMENT

BY JONATHAN EMBLETON\textsuperscript{1,a}, MARINA I. KNIGHT\textsuperscript{1,b} AND HERNANDO OMBAO\textsuperscript{2,c}

\textsuperscript{1}Department of Mathematics, University of York, \textsuperscript{2}Statistics Program, King Abdullah University of Science and Technology (KAUST), \textsuperscript{3}je687@york.ac.uk, \textsuperscript{b}marina.knight@york.ac.uk

Within the neurosciences it is natural to observe variability across time in the dynamics of an underlying brain process. Wavelets are essential in analysing brain signals because, even within a single trial, brain signals exhibit nonstationary behaviour. However, neurological signals generated within an experiment may also potentially exhibit evolution across trials (replicates), even for identical stimuli. As neurologists consider localised spectra of brain signals to be most informative, we propose the MULTiple-Trials Locally Stationary Wavelet process (MULT-LSW) that fills the gap in the literature by directly giving a stochastic wavelet representation of the time series of ordered replicates itself. MULT-LSW yields a natural desired time- and trial-localisation of the process dynamics, capturing nonstationary behaviour both within and across trials. While current techniques are restricted by the assumption of uncorrelated replicates, here we account for between-trial correlation. We rigorously develop the associated wavelet spectral estimation framework along with its asymptotic properties. By means of thorough simulation studies, we demonstrate the theoretical estimator properties hold in practice. A real data investigation into the evolutionary dynamics of the hippocampus and nucleus accumbens, during an associative learning experiment, demonstrates the applicability of our proposed methodology as well as the new insights it provides. Our model is general and facilitates wider experimental data analysis than the current literature allows.

REFERENCES


Key words and phrases. Neuroscience, wavelet-based spectra, cross-trial dependence.


BAYESIAN BI-CLUSTERING METHODS WITH APPLICATIONS IN COMPUTATIONAL BIOLOGY

BY HAN YAN¹,a, JIEXING WU²,c, YANG LI³,d AND JUN S. LIU¹,b

¹Department of Statistics, Harvard University, ahan_yan@g.harvard.edu,
bjliu@stat.harvard.edu
²Google, cwujiexingharvard@gmail.com
³Citadel Securities, dyangli.stat@gmail.com

Bi-clustering is a useful approach in analyzing large biological data sets when the observations come from heterogeneous groups and have a large number of features. We outline a general Bayesian approach in tackling bi-clustering problems in moderate to high dimensions and propose three Bayesian bi-clustering models on categorical data which increase in complexities in their modeling of the distributions of features across bi-clusters. Our proposed methods apply to a wide range of scenarios: from situations where data are cluster-distinguishable only among a small subset of features but masked by a large amount of noise to situations where different groups of data are identified by different sets of features or data exhibit hierarchical structures. Through simulation studies we show that our methods outperform existing (bi-)clustering methods in both identifying clusters and recovering feature distributional patterns across bi-clusters. We further apply the developed approaches to a human genetic dataset, a human single-cell genomic dataset, and a collection of 1774 mouse genomic datasets with a focus on 58 genes from two pathways.

REFERENCES


Key words and phrases. Bi-clustering, clustering, variable selection, categorical data, model selection, high dimensionality, genetics.


BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE
STATISTICAL
INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE