Articles

Modelling multilevel spatial behaviour in binary-mark muscle fibre configurations .................. TILMAN M. DAVIES, MATTHEW R. SCHOFIELD, JON CORNWALL AND PHILIP W. SHEARD 1329

Identifying and estimating principal causal effects in a multi-site trial of Early College High Schools .............. LO-HUA YUAN, AVI FELLER AND LUKE W. MIRATRIX 1348

Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines LIN LIU, YUQI QIU, LOKI NATARAJAN AND KAREN MESSER 1370

A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores ........ LÉKHA PATEL, NILS GUSTAFSSON, YU LIN, RAIMUND OBER, RICARDO HENRIQUES AND EDWARD COHEN 1397

Identifying multiple changes for a functional data sequence with application to freeway traffic segmentation ........ JENG-MIN CHIOU, YU-TING CHEN AND TAILEN HSING 1430

The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies .... JOANN GAGNON-BARTSCH AND YOTAM SHEM-TOV 1464

Spatio-temporal short-term wind forecast: A calibrated regime-switching method AHMED AZIZ EZZAT, MIKYOUNG JUN AND YU DING 1484

Network modelling of topological domains using Hi-C data ........ Y. X. RACHEL WANG, PURNAMRITA SARKAR, OANA URSU, ANSUL KUNDAJE AND PETER J. BICKEL 1511

Fast dynamic nonparametric distribution tracking in electron microscopic data YANJUN QIAN, JIEJIE ZHANG AND YU DING 1537

Distributional regression forests for probabilistic precipitation forecasting in complex terrain ................. LISA SCHLOSSER, TORSTEN HOTHORN, RETO STAUFFER AND ACHIM ZEILEIS 1564

Modeling seasonality and serial dependence of electricity price curves with warping functional autoregressive dynamics ............... YING CHEN, J. S. MARRON AND JIEJIE ZHANG 1590

RCRnorm: An integrated system of random-coefficient hierarchical regression models for normalizing Nanostring nCounter data .............. GAOXIANG JIA, XINLEI WANG, QIWEI LI, WEI LU, XIMING TANG, IGNACIO WISTUBA AND YANG XIE 1617

Network classification with applications to brain connectomics JESÚS D. ARROYO RELIÓN, DANIEL KESSLER, ELIZAVETA LEVINA AND STEPHAN F. TAYLOR 1648

Sequential decision model for inference and prediction on nonuniform hypergraphs with application to knot matching from computational forestry .......... SEONG-HWAN JUN, SAMUEL W. K. WONG, JAMES V. ZIDEK AND ALEXANDRE BOUCHARD-CÔTÉ 1678

A Bayesian mark interaction model for analysis of tumor pathology images QIWEI LI, XINLEI WANG, FAMING LIANG AND GUANGHUA XIAO 1708

A hierarchical Bayesian model for single-cell clustering using RNA-sequencing data YIYI LIU, JOSHUA L. WARREN AND HONGYU ZHAO 1733
Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter?...HUIPING XU, XIAOCHUN LI, CHANGYU SHEN, SIU L. HUI AND SHAUN GRANNIS 1753

Bayesian modeling of the structural connectome for studying Alzheimer’s disease........ARKAPRAVA ROY, SUBHASHIS GHOSAL, JEFFREY PRESCOTT AND KINGSHUK ROY CHOUHDURY 1791

Wavelet spectral testing: Application to nonstationary circadian rhythms
JESSICA K. HARGREAVES, MARINA I. KNIGHT, JON W. PITCHFORD, RACHAEL J. OAKENFULL, SANGEETA CHAWLA, JACK MUNNS AND Seth J. Davis 1817

Oblique random survival forests...............BYRON C. JAEGER, D. LEANN LONG, DUSTIN M. LONG, MARIO SIMS, JEFF M. SZYCHOWSKI, YUAN-I MIN, LESLIE A. McCLURE, GEORGE HOWARD AND NOAH SIMON 1847

Approximate inference for constructing astronomical catalogs from images
JEFFREY REGIER, ANDREW C. MILLER, DAVID SCHLEGEL, RYAN P. ADAMS, JON D. MCAULIFFE AND PRABHAT 1884

Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls
CHANMIN KIM, MICHAEL J. DANIELS, JOSEPH W. HOGAN, CHRISTINE CHOIRAT AND CORWIN M. ZIGLER 1927

Radio-iBAG: Radiomics-based integrative Bayesian analysis of mutiplatform genomic data........YOUYI ZHANG, JEFFREY S. MORRIS, SHIVALI NARANG AERRY, ARVIND U. K. RAO AND VEERABHADRAN BALADANDAYUTHAPANI 1957

A semiparametric modeling approach using Bayesian Additive Regression Trees with an application to evaluate heterogeneous treatment effects
BRET ZELDOW, VINCENT LO RE III AND JASON ROY 1989
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, Heidelberg Institute for Theoretical Studies, 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: Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

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 13, Number 3, September 2019. 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 © 2019 by the Institute of Mathematical Statistics
Printed in the United States of America
MODELLING MULTILEVEL SPATIAL BEHAVIOUR IN BINARY-MARK MUSCLE FIBRE CONFIGURATIONS

BY TILMAN M. DAVIES, MATTHEW R. SCHOFIELD, JON CORNWALL AND PHILIP W. SHEARD

University of Otago

The functional properties of skeletal muscles depend on the spatial arrangements of fast and slow muscle fibre types. Qualitative assessment of muscle configurations suggests that muscle disease and normal ageing are associated with visible changes in the spatial pattern, though a lack of statistical modelling hinders our ability to formally assess such trends. We design a nested Gaussian conditional autoregressive (CAR) model to quantify spatial features of dichotomously marked muscle fibre networks and implement it within a Bayesian framework. Our model is applied to data from a human skeletal muscle and results reveal spatial variation at multiple levels across the muscle. The model provides the foundation for future research in describing the extent of change to normal muscle fibre type parameters under experimental or pathological conditions.

REFERENCES


Key words and phrases. Gaussian process, hierarchical model, Bayesian inference, physiology.


HODGES, J. S. and REICH, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. Amer. Statist. 64 325–334. MR2758564


IDENTIFYING AND ESTIMATING PRINCIPAL CAUSAL EFFECTS IN A MULTI-SITE TRIAL OF EARLY COLLEGE HIGH SCHOOLS

BY LO-HUA YUAN, AVI FELLER, AND LUKE W. MIRATRIX
Airbnb, Inc., University of California, Berkeley and Harvard University

Randomized trials are often conducted with separate randomizations across multiple sites such as schools, voting districts, or hospitals. These sites can differ in important ways, including the site’s implementation quality, local conditions, and the composition of individuals. An important question in practice is whether—and under what assumptions—researchers can leverage this cross-site variation to learn more about the intervention. We address these questions in the principal stratification framework, which describes causal effects for subgroups defined by post-treatment quantities. We show that researchers can estimate certain principal causal effects via the multi-site design if they are willing to impose the strong assumption that the site-specific effects are independent of the site-specific distribution of stratum membership. We motivate this approach with a multi-site trial of the Early College High School Initiative, a unique secondary education program with the goal of increasing high school graduation rates and college enrollment. Our analyses corroborate previous studies suggesting that the initiative had positive effects for students who would have otherwise attended a low-quality high school, although power is limited.

REFERENCES


Key words and phrases. Principal causal effects, principal stratification, covariate restrictions, multi-site randomized trials, noncompliance, Early College High School.


IMPUTATION AND POST-SELECTION INFEREN CE IN MODELS WITH MISSING DATA: AN APPLICATION TO COLORECTAL CANCER SURVEILLANCE GUIDELINES

BY LIN LIU, YUQI QIU, LOKI NATARAJAN AND KAREN MESSER

University of California, San Diego

It is common to encounter missing data among the potential predictor variables in the setting of model selection. For example, in a recent study we attempted to improve the US guidelines for risk stratification after screening colonoscopy (Cancer Causes Control 27 (2016) 1175–1185), with the aim to help reduce both overuse and underuse of follow-on surveillance colonoscopy. The goal was to incorporate selected additional informative variables into a neoplasia risk-prediction model, going beyond the three currently established risk factors, using a large dataset pooled from seven different prospective studies in North America. Unfortunately, not all candidate variables were collected in all studies, so that one or more important potential predictors were missing on over half of the subjects. Thus, while variable selection was a main focus of the study, it was necessary to address the substantial amount of missing data. Multiple imputation can effectively address missing data, and there are also good approaches to incorporate the variable selection process into model-based confidence intervals. However, there is not consensus on appropriate methods of inference which address both issues simultaneously. Our goal here is to study the properties of model-based confidence intervals in the setting of imputation for missing data followed by variable selection. We use both simulation and theory to compare three approaches to such post-imputation-selection inference: a multiple-imputation approach based on Rubin’s Rules for variance estimation (Comput. Statist. Data Anal. 71 (2014) 758–770); a single imputation-selection followed by bootstrap percentile confidence intervals; and a new bootstrap model-averaging approach presented here, following Efron (J. Amer. Statist. Assoc. 109 (2014) 991–1007). We investigate relative strengths and weaknesses of each method. The “Rubin’s Rules” multiple imputation estimator can have severe undercoverage, and is not recommended. The imputation-selection estimator with bootstrap percentile confidence intervals works well. The bootstrap-model-averaged estimator, with the “Efron’s Rules” estimated variance, may be preferred if the true effect sizes are moderate. We apply these results to the colorectal neoplasia risk-prediction problem which motivated the present work.

Key words and phrases. Post-selection inference, missing data, multiple imputation, model selection, model averaging.


A HIDDEN MARKOV MODEL APPROACH TO CHARACTERIZING THE PHOTO-SWITCHING BEHAVIOR OF FLUOROPHORES

BY LEKHA PATEL*,1, NILS GUSTAFSSON†,2, YU LIN‡, RAIMUND OBER§,¶,3, RICARDO HENRIQUES†,∥,4 AND EDWARD COHEN∗,5

Imperial College London∗, University College London†, European Molecular Biology Laboratory Heidelberg‡, Texas A&M§, University of Southampton¶ and Francis Crick Institute∥

Fluorescing molecules (fluorophores) that stochastically switch between photon-emitting and dark states underpin some of the most celebrated advancements in super-resolution microscopy. While this stochastic behavior has been heavily exploited, full characterization of the underlying models can potentially drive forward further imaging methodologies. Under the assumption that fluorophores move between fluorescing and dark states as continuous time Markov processes, the goal is to use a sequence of images to select a model and estimate the transition rates. We use a hidden Markov model to relate the observed discrete time signal to the hidden continuous time process. With imaging involving several repeat exposures of the fluorophore, we show the observed signal depends on both the current and past states of the hidden process, producing emission probabilities that depend on the transition rate parameters to be estimated. To tackle this unusual coupling of the transition and emission probabilities, we conceive transmission (transition-emission) matrices that capture all dependencies of the model. We provide a scheme of computing these matrices and adapt the forward-backward algorithm to compute a likelihood which is readily optimized to provide rate estimates. When confronted with several model proposals, combining this procedure with the Bayesian Information Criterion provides accurate model selection.

REFERENCES


Key words and phrases. Hidden Markov models, Markov processes, rate estimation, forward-backward algorithm, super-resolution microscopy.


IDENTIFYING MULTIPLE CHANGES FOR A FUNCTIONAL DATA SEQUENCE WITH APPLICATION TO FREEWAY TRAFFIC SEGMENTATION

BY JENG-MIN CHIOU*,†, YU-TING CHEN‡ AND TAILEN HSING§

Academia Sinica*, National Cheng Chi University† and University of Michigan‡§

Motivated by the study of road segmentation partitioned by shifts in traffic conditions along a freeway, we introduce a two-stage procedure, Dynamic Segmentation and Backward Elimination (DSBE), for identifying multiple changes in the mean functions for a sequence of functional data. The Dynamic Segmentation procedure searches for all possible changepoints using the derived global optimality criterion coupled with the local strategy of at-most-one-changepoint by dividing the entire sequence into individual sub-sequences that are recursively adjusted until convergence. Then, the Backward Elimination procedure verifies these changepoints by iteratively testing the unlikely changes to ensure their significance until no more changepoints can be removed. By combining the local strategy with the global optimal changepoint criterion, the DSBE algorithm is conceptually simple and easy to implement and performs better than the binary segmentation-based approach at detecting small multiple changes. The consistency property of the changepoint estimators and the convergence of the algorithm are proved. We apply DSBE to detect changes in traffic streams through real freeway traffic data. The practical performance of DSBE is also investigated through intensive simulation studies for various scenarios.

REFERENCES


Key words and phrases. Changepoint analysis, covariance operator, functional principal component, projection, segmentation.


THE CLASSIFICATION PERMUTATION TEST: A FLEXIBLE APPROACH TO TESTING FOR COVARIATE IMBALANCE IN OBSERVATIONAL STUDIES

BY JOHANN GAGNON-BARTSCH AND YOTAM SHEM-TOV

University of Michigan and University of California, Berkeley

The gold standard for identifying causal relationships is a randomized controlled experiment. In many applications in the social sciences and medicine, the researcher does not control the assignment mechanism and instead may rely upon natural experiments or matching methods as a substitute to experimental randomization. The standard testable implication of random assignment is covariate balance between the treated and control units. Covariate balance is commonly used to validate the claim of as good as random assignment. We propose a new nonparametric test of covariate balance. Our Classification Permutation Test (CPT) is based on a combination of classification methods (e.g., random forests) with Fisherian permutation inference. We revisit four real data examples and present Monte Carlo power simulations to demonstrate the applicability of the CPT relative to other nonparametric tests of equality of multivariate distributions.

REFERENCES


Key words and phrases. Balance, matching, observational study, natural experiment.


SPATIO-TEMPORAL SHORT-TERM WIND FORECAST:
A CALIBRATED REGIME-SWITCHING METHOD

BY AHMED AZIZ EZZAT, MIKYOUNG JUN AND YU DING

Texas A&M University

Accurate short-term forecasts are indispensable for the integration of wind energy in power grids. On a wind farm, local wind conditions exhibit sizeable variations at a fine temporal resolution. Existing statistical models may capture the in-sample variations in wind behavior, but are often short-sighted to those occurring in the near future, that is, in the forecast horizon. The calibrated regime-switching method proposed in this paper introduces an action of regime dependent calibration on the predictand (here the wind speed variable), which helps correct the bias resulting from out-of-sample variations in wind behavior. This is achieved by modeling the calibration as a function of two elements: the wind regime at the time of the forecast (and the calibration is therefore regime dependent), and the runlength, which is the time elapsed since the last observed regime change. In addition to regime-switching dynamics, the proposed model also accounts for other features of wind fields: spatio-temporal dependencies, transport effect of wind and non-stationarity. Using one year of turbine-specific wind data, we show that the calibrated regime-switching method can offer a wide margin of improvement over existing forecasting methods in terms of both wind speed and power.

REFERENCES


Key words and phrases. Regime-switching, spatio-temporal, wind energy, wind forecast.


NETWORK MODELLING OF TOPOLOGICAL DOMAINS USING HI-C DATA

BY Y. X. RACHEL WANG*,†, PURNAMRITA SARKAR‡, OANA URSU§,2, ANSHUL KUNDaje‡ AND PETER J. BICKEL§,3

University of Sydney*, University of Texas†, Stanford University‡ and University of California, Berkeley§

Chromosome conformation capture experiments such as Hi-C are used to map the three-dimensional spatial organization of genomes. One specific feature of the 3D organization is known as topologically associating domains (TADs), which are densely interacting, contiguous chromatin regions playing important roles in regulating gene expression. A few algorithms have been proposed to detect TADs. In particular, the structure of Hi-C data naturally inspires application of community detection methods. However, one of the drawbacks of community detection is that most methods take exchangeability of the nodes in the network for granted; whereas the nodes in this case, that is, the positions on the chromosomes, are not exchangeable. We propose a network model for detecting TADs using Hi-C data that takes into account this nonexchangeability. In addition, our model explicitly makes use of cell-type specific CTCF binding sites as biological covariates and can be used to identify conserved TADs across multiple cell types. The model leads to a likelihood objective that can be efficiently optimized via relaxation. We also prove that when suitably initialized, this model finds the underlying TAD structure with high probability. Using simulated data, we show the advantages of our method and the caveats of popular community detection methods, such as spectral clustering, in this application. Applying our method to real Hi-C data, we demonstrate the domains identified have desirable epigenetic features and compare them across different cell types.

REFERENCES


Key words and phrases. Hi-C data, topologically associating domains, network models, community detection.


FAST DYNAMIC NONPARAMETRIC DISTRIBUTION TRACKING IN ELECTRON MICROSCOPIC DATA

BY YANJUN QIAN*, JIANHUA Z. HUANG†, CHIWOOL PARK‡ AND YU DING‡

Virginia Commonwealth University*, Texas A & M University† and Florida State University‡

In situ transmission electron microscope (TEM) adds a promising instrument to the exploration of the nanoscale world, allowing motion pictures to be taken while nano objects are initiating, crystallizing and morphing into different sizes and shapes. To enable in-process control of nanocrystal production, this technology innovation hinges upon a solution addressing a statistical problem, which is the capability of online tracking a dynamic, time-varying probability distribution reflecting the nanocrystal growth. Because no known parametric density functions can adequately describe the evolving distribution, a nonparametric approach is inevitable. Towards this objective, we propose to incorporate the dynamic evolution of the normalized particle size distribution into a state space model, in which the density function is represented by a linear combination of B-splines and the spline coefficients are treated as states. The closed-form algorithm runs online updates faster than the frame rate of the in situ TEM video, making it suitable for in-process control purpose. Imposing the constraints of curve smoothness and temporal continuity improves the accuracy and robustness while tracking the probability distribution. We test our method on three published TEM videos. For all of them, the proposed method is able to outperform several alternative approaches.

REFERENCES


Key words and phrases. Kalman filter, nanotechnology, nonparametric density, online density estimation, state space model.


DISTRIBUTIONAL REGRESSION FORESTS FOR PROBABILISTIC PRECIPITATION FORECASTING IN COMPLEX TERRAIN

BY LISA SCHLOSSER*, TORSTEN HOTTHORN†,1, RETO STAUFFER* AND ACHIM ZEILEIS*

Universität Innsbruck* and Universität Zürich†

To obtain a probabilistic model for a dependent variable based on some set of explanatory variables, a distributional approach is often adopted where the parameters of the distribution are linked to regressors. In many classical models this only captures the location of the distribution but over the last decade there has been increasing interest in distributional regression approaches modeling all parameters including location, scale and shape. Notably, so-called nonhomogeneous Gaussian regression (NGR) models both mean and variance of a Gaussian response and is particularly popular in weather forecasting. Moreover, generalized additive models for location, scale and shape (GAMLSS) provide a framework where each distribution parameter is modeled separately capturing smooth linear or nonlinear effects. However, when variable selection is required and/or there are nonsmooth dependencies or interactions (especially unknown or of high-order), it is challenging to establish a good GAMLSS. A natural alternative in these situations would be the application of regression trees or random forests but, so far, no general distributional framework is available for these. Therefore, a framework for distributional regression trees and forests is proposed that blends regression trees and random forests with classical distributions from the GAMLSS framework as well as their censored or truncated counterparts. To illustrate these novel approaches in practice, they are employed to obtain probabilistic precipitation forecasts at numerous sites in a mountainous region (Tyrol, Austria) based on a large number of numerical weather prediction quantities. It is shown that the novel distributional regression forests automatically select variables and interactions, performing on par or often even better than GAMLSS specified either through prior meteorological knowledge or a computationally more demanding boosting approach.

REFERENCES


Key words and phrases. Parametric models, regression trees, random forests, recursive partitioning, probabilistic forecasting, GAMLSS.


MODELING SEASONALITY AND SERIAL DEPENDENCE OF ELECTRICITY PRICE CURVES WITH WARPING FUNCTIONAL AUTOREGRESSIVE DYNAMICS

BY YING CHEN* J. S. MARRON*,† 2 AND JIEJIE ZHANG*

National University of Singapore* and University of North Carolina†

Electricity prices are high dimensional, serially dependent and have seasonal variations. We propose a Warping Functional AutoRegressive (WFAR) model that simultaneously accounts for the cross time-dependence and seasonal variations of the large dimensional data. In particular, electricity price curves are obtained by smoothing over the 24 discrete hourly prices on each day. In the functional domain, seasonal phase variations are separated from level amplitude changes in a warping process with the Fisher–Rao distance metric, and the aligned (season-adjusted) electricity price curves are modeled in the functional autoregression framework. In a real application, the WFAR model provides superior out-of-sample forecast accuracy in both a normal functioning market, Nord Pool, and an extreme situation, the California market. The forecast performance as well as the relative accuracy improvement are stable for different markets and different time periods.

REFERENCES


Key words and phrases. Seasonal functional time series, warping function, Karcher mean.


RCRNORM: AN INTEGRATED SYSTEM OF RANDOM-COEFFICIENT HIERARCHICAL REGRESSION MODELS FOR NORMALIZING NANOSTRING NCOUNTER DATA

BY GAOXIANG JIA*,†, XINLEI WANG*, QIWEI LI‡, WEI LU§, XIMING TANG§, IGNACIO WISTUBA§ AND YANG XIE†

Southern Methodist University*, University of Texas Southwestern Medical Center‡, University of Texas at Dallas§ and University of Texas§

Formalin-fixed paraffin-embedded (FFPE) samples have great potential for biomarker discovery, retrospective studies and diagnosis or prognosis of diseases. Their application, however, is hindered by the unsatisfactory performance of traditional gene expression profiling techniques on damaged RNAs. NanoString nCounter platform is well suited for profiling of FFPE samples and measures gene expression with high sensitivity which may greatly facilitate realization of scientific and clinical values of FFPE samples. However, methodological development for normalization, a critical step when analyzing this type of data, is far behind. Existing methods designed for the platform use information from different types of internal controls separately and rely on an overly-simplified assumption that expression of housekeeping genes is constant across samples for global scaling. Thus, these methods are not optimized for the nCounter system, not mentioning that they were not developed for FFPE samples. We construct an integrated system of random-coefficient hierarchical regression models to capture main patterns and characteristics observed from NanoString data of FFPE samples and develop a Bayesian approach to estimate parameters and normalize gene expression across samples. Our method, labeled RCRnorm, incorporates information from all aspects of the experimental design and simultaneously removes biases from various sources. It eliminates the unrealistic assumption on housekeeping genes and offers great interpretability. Furthermore, it is applicable to freshly frozen or like samples that can be generally viewed as a reduced case of FFPE samples. Simulation and applications showed the superior performance of RCRnorm.

REFERENCES


Key words and phrases. Bayesian hierarchical modeling, control probes, FFPE, housekeeping gene, normalization, random coefficients regression.


NETWORK CLASSIFICATION WITH APPLICATIONS TO BRAIN CONNECTOMICS

BY JESÚS D. ARROYO RELIÓN*, DANIEL KESSLER†, ELIZAVETA LEVINA†
AND STEPHAN F. TAYLOR†,2

Johns Hopkins University* and University of Michigan†

While statistical analysis of a single network has received a lot of attention in recent years, with a focus on social networks, analysis of a sample of networks presents its own challenges which require a different set of analytic tools. Here we study the problem of classification of networks with labeled nodes, motivated by applications in neuroimaging. Brain networks are constructed from imaging data to represent functional connectivity between regions of the brain, and previous work has shown the potential of such networks to distinguish between various brain disorders, giving rise to a network classification problem. Existing approaches tend to either treat all edge weights as a long vector, ignoring the network structure, or focus on graph topology as represented by summary measures while ignoring the edge weights. Our goal is to design a classification method that uses both the individual edge information and the network structure of the data in a computationally efficient way, and that can produce a parsimonious and interpretable representation of differences in brain connectivity patterns between classes. We propose a graph classification method that uses edge weights as predictors but incorporates the network nature of the data via penalties that promote sparsity in the number of nodes, in addition to the usual sparsity penalties that encourage selection of edges. We implement the method via efficient convex optimization and provide a detailed analysis of data from two fMRI studies of schizophrenia.

REFERENCES


Key words and phrases. Graph classification, high-dimensional data, variable selection, fMRI data.


SEQUENTIAL DECISION MODEL FOR INference AND PREDICTION ON NONUNIFORM HYPERGRAPHS WITH APPLICATION TO KNOT MATCHING FROM COMPUTATIONAL FORESTRY

BY SEONG-HWAN JUN*, SAMUEL W. K. WONG†, JAMES V. ZIDK* AND ALEXANDRE BOUCHARD-CÔTÉ*

University of British Columbia* and University of Waterloo†

In this paper, we consider the knot-matching problem arising in computational forestry. The knot-matching problem is an important problem that needs to be solved to advance the state of the art in automatic strength prediction of lumber. We show that this problem can be formulated as a quadripartite matching problem and develop a sequential decision model that admits efficient parameter estimation along with a sequential Monte Carlo sampler on graph matching that can be utilized for rapid sampling of graph matching. We demonstrate the effectiveness of our methods on 30 manually annotated boards and present findings from various simulation studies to provide further evidence supporting the efficacy of our methods.

REFERENCES


Key words and phrases. Graph matching, sequential Monte Carlo, expectation maximization, Plackett–Luce model, computational forestry.


A BAYESIAN MARK INTERACTION MODEL FOR ANALYSIS OF TUMOR PATHOLOGY IMAGES

BY QIWEI LI*, XINLEI WANG†, FAMING LIANG‡ AND GUANGHUA XIAO§

University of Texas at Dallas*, Southern Methodist University†, Purdue University‡ and University of Texas Southwestern Medical Center§

With the advance of imaging technology, digital pathology imaging of tumor tissue slides is becoming a routine clinical procedure for cancer diagnosis. This process produces massive imaging data that capture histological details in high resolution. Recent developments in deep-learning methods have enabled us to identify and classify individual cells from digital pathology images at large scale. Reliable statistical approaches to model the spatial pattern of cells can provide new insight into tumor progression and shed light on the biological mechanisms of cancer. We consider the problem of modeling spatial correlations among three commonly seen cells observed in tumor pathology images. A novel geostatistical marking model with interpretable underlying parameters is proposed in a Bayesian framework. We use auxiliary variable MCMC algorithms to sample from the posterior distribution with an intractable normalizing constant. We demonstrate how this model-based analysis can lead to sharper inferences than ordinary exploratory analyses, by means of application to three benchmark datasets and a case study on the pathology images of 188 lung cancer patients. The case study shows that the spatial correlation between tumor and stromal cells predicts patient prognosis. This statistical methodology not only presents a new model for characterizing spatial correlations in a multitype spatial point pattern conditioning on the locations of the points, but also provides a new perspective for understanding the role of cell–cell interactions in cancer progression.

REFERENCES


Key words and phrases. Multitype point pattern, spatial correlation, Markov random field, double Metropolis–Hastings.


A HIERARCHICAL BAYESIAN MODEL FOR SINGLE-CELL CLUSTERING USING RNA-SEQUENCING DATA

BY YIYI LIU¹, JOSHUA L. WARREN² AND HONGYU ZHAO¹

Yale University

Understanding the heterogeneity of cells is an important biological question. The development of single-cell RNA-sequencing (scRNA-seq) technology provides high resolution data for such inquiry. A key challenge in scRNA-seq analysis is the high variability of measured RNA expression levels and frequent dropouts (missing values) due to limited input RNA compared to bulk RNA-seq measurement. Existing clustering methods do not perform well for these noisy and zero-inflated scRNA-seq data. In this manuscript we propose a Bayesian hierarchical model, called BasClu, to appropriately characterize important features of scRNA-seq data in order to more accurately cluster cells. We demonstrate the effectiveness of our method with extensive simulation studies and applications to three real scRNA-seq datasets.

REFERENCES


Key words and phrases. Bayesian hierarchical model, clustering, Dirichlet process, Gaussian mixture model, missing data, single-cell RNA-sequencing.


INCORPORATING CONDITIONAL DEPENDENCE IN LATENT CLASS MODELS FOR PROBABILISTIC RECORD LINKAGE: DOES IT MATTER?¹

BY HUIPING XU*, XIAOCHUN LI*, CHANGYU SHEN†, SIU L. HUI‡ AND SHAUN GRANNIS*,‡

Indiana University*, Harvard Medical School† and Regenstrief Institute‡

The conditional independence assumption of the Felligi and Sunter (FS) model in probabilistic record linkage is often violated when matching real-world data. Ignoring conditional dependence has been shown to seriously bias parameter estimates. However, in record linkage, the ultimate goal is to inform the match status of record pairs and therefore, record linkage algorithms should be evaluated in terms of matching accuracy. In the literature, more flexible models have been proposed to relax the conditional independence assumption, but few studies have assessed whether such accommodations improve matching accuracy. In this paper, we show that incorporating the conditional dependence appropriately yields comparable or improved matching accuracy than the FS model using three real-world data linkage examples. Through a simulation study, we further investigate when conditional dependence models provide improved matching accuracy. Our study shows that the FS model is generally robust to the conditional independence assumption and provides comparable matching accuracy as the more complex conditional dependence models. However, when the match prevalence approaches 0% or 100% and conditional dependence exists in the dominating class, it is necessary to address conditional dependence as the FS model produces suboptimal matching accuracy. The need to address conditional dependence becomes less important when highly discriminating fields are used. Our simulation study also shows that conditional dependence models with misspecified dependence structure could produce less accurate record matching than the FS model and therefore we caution against the blind use of conditional dependence models.

REFERENCES


Key words and phrases. Conditional dependence, finite mixture, Gaussian random effects model, latent class analysis, log-linear model, record linkage.


BAYESIAN MODELING OF THE STRUCTURAL CONNECTOME FOR STUDYING ALZHEIMER’S DISEASE

BY ARKAPRAVA ROY*,1, SUBHASHIS GHOSAL†,1, JEFFREY PRESCOTT‡, KINGSHUK ROY CHOUdhury* AND FOR THE ALZHEIMER’S DISEASE NEUROIMAGING INITIATIVE2

Duke University*, North Carolina State University† and MetroHealth Medical Center‡

We study possible relations between Alzheimer’s disease progression and the structure of the connectome which is white matter connecting different regions of the brain. Regression models in covariates including age, gender and disease status for the extent of white matter connecting each pair of regions of the brain are proposed. Subject inhomogeneity is also incorporated in the model through random effects with an unknown distribution. As there is a large number of pairs of regions, we also adopt a dimension reduction technique through graphon (J. Combin. Theory Ser. B 96 (2006) 933–957) functions which reduces the functions of pairs of regions to functions of regions. The connecting graphon functions are considered unknown but the assumed smoothness allows putting priors of low complexity on these functions. We pursue a nonparametric Bayesian approach by assigning a Dirichlet process scale mixture of zero to mean normal prior on the distributions of the random effects and finite random series of tensor products of B-splines priors on the underlying graphon functions. We develop efficient Markov chain Monte Carlo techniques for drawing samples for the posterior distributions using Hamiltonian Monte Carlo (HMC). The proposed Bayesian method overwhelmingly outperforms a competing method based on ANCOVA models in the simulation setup. The proposed Bayesian approach is applied on a dataset of 100 subjects and 83 brain regions and key regions implicated in the changing connectome are identified.

REFERENCES


Key words and phrases. ADNI, B-spline prior, brain image, connectome, graphical model, Graphon, HMC.


WAVELET SPECTRAL TESTING: APPLICATION TO NONSTATIONARY CIRCADIAN RHYTHMS

BY JESSICA K. HARGREAVES, MARINA I. KNIGHT, JON W. PITCHFORD, RACHAEL J. OAKENFULL, SANGEETA CHAWLA, JACK MUNNS AND SETH J. DAVIS

University of York

Rhythmic data are ubiquitous in the life sciences. Biologists need reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in a rhythmic signal. When these signals display nonstationary behaviour, as is common in many biological systems, the established methodologies may be misleading. Therefore, there is a real need for new methodology that enables the formal comparison of nonstationary processes. As circadian behaviour is best understood in the spectral domain, here we develop novel hypothesis testing procedures in the (wavelet) spectral domain, embedding replicate information when available. The data are modelled as realisations of locally stationary wavelet processes, allowing us to define and rigorously estimate their evolutionary wavelet spectra. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. We demonstrate the advantages of our methodology over alternative approaches, by means of a comprehensive simulation study and real data applications, using both published and newly generated circadian datasets. In contrast to the current standard methodologies, our method successfully identifies differences within the motivating circadian datasets, and facilitates wider ranging analyses of rhythmic biological data in general.

REFERENCES


Key words and phrases. Wavelets, spectral decomposition, hypothesis testing, circadian rhythms.


OBLIQUE RANDOM SURVIVAL FORESTS

BY BYRON C. JAEGERT®, D. LEANN LONG®, DUSTIN M. LONG®,
MARIO SIMSTM, JEFF M. SZYCHOWSKIT®, YUAN-I MINTM,
LESLIE A. MCCULURE‡, GEORGE HOWARDT® AND NOAH SIMONT®

University of Alabama at Birmingham®, University of Mississippi Medical
Center†, Dornsife School of Public Health Drexel University‡ and
University of Washington§

We introduce and evaluate the oblique random survival forest (ORSF). The ORSF is an ensemble method for right-censored survival data that uses linear combinations of input variables to recursively partition a set of training data. Regularized Cox proportional hazard models are used to identify linear combinations of input variables in each recursive partitioning step. Benchmark results using simulated and real data indicate that the ORSF’s predicted risk function has high prognostic value in comparison to random survival forests, conditional inference forests, regression and boosting. In an application to data from the Jackson Heart Study, we demonstrate variable and partial dependence using the ORSF and highlight characteristics of its ten-year predicted risk function for atherosclerotic cardiovascular disease events (ASCVD; stroke, coronary heart disease). We present visualizations comparing variable and partial effect estimation according to the ORSF, the conditional inference forest, and the Pooled Cohort Risk equations. The obliqueRSF R package, which provides functions to fit the ORSF and create variable and partial dependence plots, is available on the comprehensive R archive network (CRAN).

REFERENCES


Key words and phrases. Random forest, survival, machine learning, penalized regression, cardiovascular disease.


Schumacher, M., Baster, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. et al. (1994). Randomized 2 × 2 trial


APPROXIMATE INFERENCE FOR CONSTRUCTING ASTRONOMICAL CATALOGS FROM IMAGES

BY JEFFREY REGIER*, ANDREW C. MILLER†, DAVID SCHLEGL‡, RYAN P. ADAMS§, JON D. MCAULIFFE*,¶ AND PRABHAT‡

University of California, Berkeley*, Columbia University†, Lawrence Berkeley National Laboratory‡, Princeton University§ and The Voleon Group¶

We present a new, fully generative model for constructing astronomical catalogs from optical telescope image sets. Each pixel intensity is treated as a random variable with parameters that depend on the latent properties of stars and galaxies. These latent properties are themselves modeled as random. We compare two procedures for posterior inference. One procedure is based on Markov chain Monte Carlo (MCMC) while the other is based on variational inference (VI). The MCMC procedure excels at quantifying uncertainty, while the VI procedure is 1000 times faster. On a supercomputer, the VI procedure efficiently uses 665,000 CPU cores to construct an astronomical catalog from 50 terabytes of images in 14.6 minutes, demonstrating the scaling characteristics necessary to construct catalogs for upcoming astronomical surveys.

REFERENCES


Key words and phrases. Astronomy, graphical model, MCMC, variational inference, high performance computing.


BAYESIAN METHODS FOR MULTIPLE MEDIATORS: RELATING PRINCIPAL STRATIFICATION AND CAUSAL MEDIATION IN THE ANALYSIS OF POWER PLANT EMISSION CONTROLS

BY CHANMIN KIM∗, MICHAEL J. DANIELS†, JOSEPH W. HOGAN‡, CHRISTINE CHOIRAT§ AND CORWIN M. ZIGLER¶

Boston University School of Public Health∗, University of Florida†, Brown University School of Public Health‡, Swiss Data Science Center§, University of Texas at Austin¶

Emission control technologies installed on power plants are a key feature of many air pollution regulations in the US. While such regulations are predicated on the presumed relationships between emissions, ambient air pollution and human health, many of these relationships have never been empirically verified. The goal of this paper is to develop new statistical methods to quantify these relationships. We frame this problem as one of mediation analysis to evaluate the extent to which the effect of a particular control technology on ambient pollution is mediated through causal effects on power plant emissions. Since power plants emit various compounds that contribute to ambient pollution, we develop new methods for multiple intermediate variables that are measured contemporaneously, may interact with one another, and may exhibit joint mediating effects. Specifically, we propose new methods leveraging two related frameworks for causal inference in the presence of mediating variables: principal stratification and causal mediation analysis. We define principal effects based on multiple mediators, and also introduce a new decomposition of the total effect of an intervention on ambient pollution into the natural direct effect and natural indirect effects for all combinations of mediators. Both approaches are anchored to the same observed-data models, which we specify with Bayesian nonparametric techniques. We provide assumptions for estimating principal causal effects, then augment these with an additional assumption required for causal mediation analysis. The two analyses, interpreted in tandem, provide the first empirical investigation of the presumed causal pathways that motivate important air quality regulatory policies.

REFERENCES


Key words and phrases. Ambient PM$_{2.5}$, Bayesian nonparametrics, Gaussian copula, multipollutants, natural indirect effect.


U. S. EPA (2013). Workshop on designing research to assess air quality and health outcomes from air pollution regulations. In *Designing Research to Assess Air Quality and Health Outcomes from Air Pollution Regulations.*


RADIO-IBAG: RADIOMICS-BASED INTEGRATIVE BAYESIAN ANALYSIS OF MULTIPLATFORM GENOMIC DATA

BY YOUYI ZHANG∗, JEFFREY S. MORRIS∗, SHIVALI NARANG AERRY†, ARVIND U. K. RAO‡, AND VEERABHADRAN BALADANDAYUTHAPANI‡,1

The University of Texas MD Anderson Cancer Center,∗ Johns Hopkins University† and University of Michigan‡

Technological innovations have produced large multi-modal datasets that include imaging and multi-platform genomics data. Integrative analyses of such data have the potential to reveal important biological and clinical insights into complex diseases like cancer. In this paper, we present Bayesian approaches for integrative analysis of radiological imaging and multi-platform genomic data, where-in our goals are to simultaneously identify genomic and radiomic, that is, radiology-based imaging markers, along with the latent associations between these two modalities, and to detect the overall prognostic relevance of the combined markers. For this task, we propose Radio-iBAG: Radiomics-based Integrative Bayesian Analysis of Multi-platform Genomic Data, a multi-scale Bayesian hierarchical model that involves several innovative strategies: it incorporates integrative analysis of multi-platform genomic data sets to capture fundamental biological relationships; explores the associations between radiomic markers accompanying genomic information with clinical outcomes; and detects genomic and radiomic markers associated with clinical prognosis. We also introduce the use of sparse Principal Component Analysis (sPCA) to extract a sparse set of approximately orthogonal meta-features each containing information from a set of related individual radiomic features, reducing dimensionality and combining like features. Our methods are motivated by and applied to The Cancer Genome Atlas glioblastoma multiforme data set, where-in we integrate magnetic resonance imaging-based biomarkers along with genomic, epigenomic and transcriptomic data. Our model identifies important magnetic resonance imaging features and the associated genomic platforms that are related with patient survival times.

REFERENCES


Key words and phrases. Radiological imaging, multi-platform genomics, Bayesian modeling, sparsity priors, integrative modeling, multi-scale models, cancer.


A SEMIPARAMETRIC MODELING APPROACH USING BAYESIAN ADDITIVE REGRESSION TREES WITH AN APPLICATION TO EVALUATE HETEROGENEOUS TREATMENT EFFECTS

BY BRET ZELDOW, VINCENT LO RE III AND JASON ROY

Harvard Medical School, Perelman School of Medicine and Rutgers School of Public Health

Bayesian Additive Regression Trees (BART) is a flexible machine learning algorithm capable of capturing nonlinearities between an outcome and covariates and interactions among covariates. We extend BART to a semiparametric regression framework in which the conditional expectation of an outcome is a function of treatment, its effect modifiers, and confounders. The confounders are allowed to have unspecified functional form, while treatment and effect modifiers that are directly related to the research question are given a linear form. The result is a Bayesian semiparametric linear regression model where the posterior distribution of the parameters of the linear part can be interpreted as in parametric Bayesian regression. This is useful in situations where a subset of the variables are of substantive interest and the others are nuisance variables that we would like to control for. An example of this occurs in causal modeling with the structural mean model (SMM). Under certain causal assumptions, our method can be used as a Bayesian SMM. Our methods are demonstrated with simulation studies and an application to dataset involving adults with HIV/Hepatitis C coinfection who newly initiate antiretroviral therapy. The methods are available in an R package called semibart.

REFERENCES


Key words and phrases. Bayesian Additive Regression Trees, structural mean model, antiretrovirals.


The Annals of Applied Statistics

Next Issues

Fitting a deeply-nested hierarchical model to a large book review dataset using a moment-based estimator ............... NINGSHAN ZHANG, KYLE SCHMAUS AND PATRICK O. PERRY
Bayesian factor models for probabilistic cause of death assessment with verbal autopsies ......................... TSUYOSHI KUNIHAMA, ZEHANG LI, SAMUEL CLARK AND TYLER MCCORMICK

A latent discrete Markov random field approach to identifying and classifying historical forest communities based on spatial multivariate tree species counts .............. STEPHEN BERG, JUN ZHU, MURRAY K. CLAYTON, MONIKA E. SHEA AND DAVID J. MLADENOFF
A nonparametric spatial test to identify factors that shape a microbiome... SUSHEELA P. SINGH, ANA-MARIA STAICU, ROBERT R. DUNN, NOAH FIERER AND BRIAN J. REICH

Estimating the rate constant from biosensor data via an adaptive variational Bayesian approach....... YE ZHANG, ZHIGANG YAO, PATRIK FORSSEN AND TORGNY FORNSTEDT
Estimating abundance from multiple sampling capture-recapture data via a multi-state multi-period stopover model ........... HANNAH WORTHINGTON, RACHEL MCCREA, RUTH KING AND RICHARD GRIFFITHS

SHOPPER: A probabilistic model of consumer choice with substitutes and complements ........................................ FRANCISCO J. R. RUIZ
A hierarchical curve-based approach to the analysis of manifold data .......... LIBERTY VITTERT, ADRIAN W. BOWMAN AND STANISLAV KATINA

BART with targeted smoothing: An analysis of patient-specific stillbirth risk
JENNIFER E. STARLING, JARED S. MURRAY, CARLOS M. CARVALHO, RADEK K. BUKOWSKI AND JAMES G. SCOTT

Robust elastic net estimators for variable selection and identification of proteomic biomarkers.......................... GABRIELA V. COHEN FREUE, DAVID KEPPLINGER, MATIAS SALIBIAN-BARRERA AND EZEQUIEL SMUCLER

Empirical Bayes analysis of RNA sequencing experiments with auxiliary information
KUN LIANG

Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications.................. LI ZHU, ZHIGUANG HUO, TIANZHOU MA, STEFFI OESTERREICH AND GEORGE C. TSENG

Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis
JASON Xu, SAMSON KOELLE, PETER GUTTORP, CHUANFENG WU, CYNTHIA DUNBAR, JANIS ABKOWITZ AND VLADIMIR MININ

Fire seasonality identification with multimodality tests
JOSE AMEJEIRAS-ALONSO, AKLI BENALI, ROSA M. CRUJEIRAS, ALBERTO RODRIGUEZ-CASAL AND JOSÉ M. C. PEREIRA

Joint model of accelerated failure time and mechanistic nonlinear model for censored covariates, with application in HIV/AIDS ....................... HONGBIN ZHANG AND LANG WU

Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways ... FEDERICO CASTELLETTI AND GUIDO CONSONNI
Prediction of small area quantiles for the conservation effects assessment project using a mixed effects quantile regression model ............... EMILY BERG AND DANHYANG LEE

Continued
The Annals of Applied Statistics

Next Issues—Continued

Principal nested shape space analysis of molecular dynamics data
IAN L. DRYDEN, KWANG-RAE KIM, CHARLES A. LAUGHTON AND HUILING LE

Efficient real-time monitoring of an emerging influenza pandemic: How feasible?
PALL JEFFREY BIRRELL, LORENZ WERNISCH, BRIAN D. M. TOM, LEONHARD HELD,
GARETH O. ROBERTS, RICHARD G. PEBODY, DANIELA DE ANGELIS

Microsimulation model calibration using incremental mixture approximate Bayesian
computation .............. CAROLYN M. RUTTER, JONATHAN OZIK, MARIA DEYOREO
AND NICHOLSON COLLIER

Spatial modeling of trends in crime over time in Philadelphia
CECILIA BALOCCHI AND SHANE JENSEN

Predicting paleoclimate from compositional data using multivariate Gaussian process inverse
prediction ................. JOHN TIPTON, MEVIN HOOTEN, CONNOR NOLAN,
ROBERT BOOTH AND JASON MCLACHLAN

Propensity score weighting for causal inference with multiple treatments . . . FAN LI AND FAN LI

Modeling microbial abundances and dysbiosis with beta-binomial regression
BRYAN D. MARTIN, DANIELA WITTEN AND AMY WILLIS

Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods:
A winning solution to the NIJ “Real-Time Crime Forecasting Challenge”
SETH FLAXMAN, MICHAEL CHIRICO, PAU PEREIRA AND CHARLES LOEFFLER

Outline analyses of the called strike zone in Major League baseball
DALE ZIMMERMAN, JUN TANG AND RUI HUANG

On Bayesian new edge prediction and anomaly detection in computer networks
SILVIA METELLI AND NICHOLAS HEARD

Integrative survival analysis with uncertain event times in application to a suicide risk
study ....................... WEIJIE WANG, ROBERT ASELTINE, KUN CHEN AND JUN YAN

A general theory for preferential sampling in environmental networks
JOE WATSON, JAMES ZIDEK AND GAVIN SHADDICK

New formulation of the Logistic-Gaussian process to analyze trajectory tracking data
GIANNLUCA MASTRANTONIO, CLARA GRAZIAN,
SARA MANCINELLI AND ENRICO BIBBONA

A statistical analysis of noisy crowdsourced weather data
ARNAB CHAKRABORTY, SOUMENDRA NATH LAHIRI AND ALYSON WILSON

A simple, consistent estimator of heritability for genome-wide association studies
ARMIN SCHWARTZMAN, ANDREW J. SCHORK, RONG ZABLOCKI
AND WESLEY K. THOMPSON

Hierarchical infinite factor models for improving the prediction of surgical complications for
geriatric patients . . . . . ELIZABETH LORENZI, RICARDO HENAO AND KATHERINE HELLER

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 MURPHY