

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

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 LEKHA 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 TAIEN HSING 1430
- The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies JOHANN 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, ANSHUL KUNDAJE AND PETER J. BICKEL 1511
- Fast dynamic nonparametric distribution tracking in electron microscopic data
YANJUN QIAN, JIANHUA Z. HUANG, CHIWO PARK 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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—Continued from front cover

- 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 CHOUDHURY 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 multiplatform 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

THE ANNALS OF APPLIED STATISTICS

Vol. 13, No. 3, pp. 1329–2010 September 2019

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

The Annals of Applied Probability. *Editors:* François Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

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.

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 suggest 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

- AARE, S., SPENDIFF, S., VUDA, M., ELKRIEF, D., PEREZ, A., WU, Q., MAYAKI, D., HUS-SAIN, S. N. A., HETTWER, S. and HEPPLE, R. T. (2016). Failed reinnervation in ageing skeletal muscle. *Skeletal Muscle* **6** 29.
- ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. [MR2461882](#)
- BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. [MR3362184](#)
- BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. [MR2523906](#)
- CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.
- CORNWALL, J. and KENNEDY, E. (2015). Fibre types of the anterior and lateral cervical muscles. *Eur. Spine J.* **24** 1986–1991.
- CORNWALL, J. and SHEARD, P. W. (2012). Spatial aggregation of type II fibres in human cervical muscles. *Clin. Anat.* **25** 533.
- CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1239641](#)
- DAVIES, T. M., CORNWALL, J. and SHEARD, P. W. (2013). Modelling dichotomously marked muscle fibre configurations. *Stat. Med.* **32** 4240–4258. [MR3118352](#)
- DAVIES, T. M., SHEARD, P. W. and CORNWALL, J. (2016). Comment on Makino et al. and observations on spatial modeling. *Anat. Sci. Int.* **91** 423–424.

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

- DAVIES, T. M., SCHOFIELD, M. R., CORNWALL, J. and SHEARD, P. W. (2019). Supplement to “Modelling multilevel spatial behaviour in binary-mark muscle fibre configurations.” DOI:10.1214/18-AOAS1214SUPPA, DOI:10.1214/18-AOAS1214SUPPB.
- DURANTE, D. (2018). Conjugate Bayes for probit regression via unified skew-normals. Available at [arXiv:1802.09565](https://arxiv.org/abs/1802.09565).
- EDELSBRUNNER, H., KIRKPATRICK, D. G. and SEIDEL, R. (1983). On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory* **29** 551–559. [MR0713690](#)
- FAULKNER, J. A., LARKIN, L. M., CLAFLIN, D. R. and BROOKS, S. V. (2007). Age-related changes in the structure and function of skeletal muscles. *Clin. Exp. Pharmacol. Physiol.* **34** 1091–1096.
- FINLEY, A. O., BANERJEE, S. and GELFAND, A. E. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *J. Stat. Softw.* **63** 1–28.
- FONSECA, S., WILSON, I. J., HORGAN, G. W. and MALTIN, C. A. (2003). Slow fiber cluster pattern in pig longissimus thoracis muscle: Implications for myogenesis. *J. Anim. Sci.* **81** 973–983.
- GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. and SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* **13** 263–312. [MR2154003](#)
- 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](#)
- JIN, X., CARLIN, B. P. and BANERJEE, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61** 950–961. [MR2216188](#)
- KELLY, N. A., HAMMOND, K. G., STEC, M. J., BICKEL, C. S., WINDHAM, S. T., TUGGLE, S. C. and BAMMAN, M. M. (2018). Quantification and characterization of grouped type I myofibers in human aging. *Muscle Nerve* **57** E52–E59.
- LEXELL, J., TAYLOR, C. C. and SJÖSTRÖM, M. (1988). What is the cause of the ageing atrophy? Total number, size and proportion of different fiber types studied in whole vastus lateralis muscle from 15- to 83-year-old men. *J. Neurol. Sci.* **84** 275–294.
- MAKINO, K., FUNAYAMA, K. and IKEGAYA, Y. (2016). Spatial clusters of constitutively active neurons in mouse visual cortex. *Anat. Sci. Int.* **91** 188–195.
- MILLER, A., WOODLEY, S. and CORNWALL, J. (2016). Fibre types of the longus capitis and longus colli muscles in elderly females. *Anat. Sci. Int.* **91** 163–168.
- MOSOLE, S., CARRARO, U., KERN, H., LOEFLER, S., FRUHMANN, H., VOGELAUER, M., BURGGRAF, S., MAYR, W., KRENN, M., PATERNOSTRO-SLUGA, T. et al. (2014). Long-term high-level exercise promotes muscle reinnervation with age. *J. Neuropathol. Exp. Neurol.* **73** 284–294.
- NEW ZEALAND GOVT. (2008). Human Tissue Act. Public Act No 28, Wellington.
- OBERBACH, A., BOSSENZ, Y., LEHMANN, S., NIEBAUER, J., ADAMS, V., PASCHKE, R., SCHÖN, M. R., BLÜHER, M. and PUNKT, K. (2006). Altered fiber distribution and fiber-specific glycolytic and oxidative enzyme activity in skeletal muscle of patients with type 2 diabetes. *Diabetes Care* **29** 895–900.
- OKABE, A., BOOTS, B., SUGIHARA, K. and CHIU, S. N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR1770006](#)
- PATEIRO-LÓPEZ, B. and RODRÍGUEZ-CASAL, A. (2010). Generalizing the convex hull of a sample: The R package alphahull. *J. Stat. Softw.* **34** 1–28.
- PERNUŠ, F. (1988). The Delaunay triangulation and the shape hull as tools in muscle fibre analysis. *Pattern Recogn. Lett.* **8** 197–202.
- PETTITT, A. N., WEIR, I. S. and HART, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. Comput.* **12** 353–367. [MR1951708](#)

- R CORE TEAM (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- REN, Q. and BANERJEE, S. (2013). Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach. *Biometrics* **69** 19–30. [MR3058048](#)
- RUEDEN, C. T., SCHINDELIN, J., HINER, M. C., DEZONIA, B. E., WALTER, A. E., ARENA, E. T. and ELICEIRI, K. W. (2017). ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinform.* **18** 529.
- TAKASHINA, N., BEGER, M., KUSUMOTO, B., RATHNAYAKE, S. and POSSINGHAM, H. P. (2018). A theory for ecological survey methods to map individual distributions. *Theor. Ecol.* **11** 213–223.
- VENEMA, H. W. (1993). Estimation of the parameters of a binary Markov random field on a graph with application to fibre type distributions in a muscle cross-section. *IMA J. Math. Appl. Med. Biol.* **10** 115–133.
- WACKERNAGEL, H. (1998). *Multivariate Geostatistics: An Introduction with Applications*, 2nd ed. Springer, New York.
- WEBSTER, C., SILBERSTEIN, L., HAYS, A. P. and BLAU, H. M. (1988). Fast muscle fibers are preferentially affected in Duchenne muscular dystrophy. *Cell* **52** 503–513.
- WEIR, I. S. and PETTITT, A. N. (1999). Spatial modelling for binary data using a hidden conditional autoregressive Gaussian process: A multivariate extension of the probit model. *Stat. Comput.* **9** 77–86.
- WEIR, I. S. and PETTITT, A. N. (2000). Binary probability maps using a hidden conditional autoregressive Gaussian process with an application to Finnish common toad data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 473–484. [MR1824553](#)

IDENTIFYING AND ESTIMATING PRINCIPAL CAUSAL EFFECTS IN A MULTI-SITE TRIAL OF EARLY COLLEGE HIGH SCHOOLS¹

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

- ALLENSWORTH, E. (2005). Graduation and Dropout Trends in Chicago: A Look at Cohorts of Students from 1991 through 2004. Report Highlights. Consortium on Chicago School Research. Available at <https://files.eric.ed.gov/fulltext/ED486035.pdf>.
- ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.
- BLOOM, H. S., RAUDENBUSH, S. W., WEISS, M. J. and PORTER, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness* **10** 817–842.
- BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525.
- CARNEGIE, N. B., HARADA, M. and HILL, J. L. (2016). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* **9** 395–420.
- DING, P., GENG, Z., YAN, W. and ZHOU, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Amer. Statist. Assoc.* **106** 1578–1591. [MR2896858](#)

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

- EDMUNDS, J. A., BERNSTEIN, L., GLENNIE, E., WILLSE, J., ARSHAVSKY, N., UNLU, F., BARTZ, D., SILBERMAN, T., SCALES, W. D. et al. (2010). Preparing students for college: The implementation and impact of the Early College High School model. *Peabody Journal of Education* **85** 348–364.
- EDMUNDS, J. A., BERNSTEIN, L., UNLU, F., GLENNIE, E., WILLSE, J., SMITH, A. and ARSHAVSKY, N. (2012). Expanding the start of the College Pipeline: Ninth-grade findings from an experimental study of the impact of the Early College High School model. *Journal of Research on Educational Effectiveness* **5** 136–159.
- EDMUNDS, J. A., UNLU, F., GLENNIE, E., BERNSTEIN, L., FESLER, L., FUREY, J. and ARSHAVSKY, N. (2017). Smoothing the transition to postsecondary education: The impact of the Early College model. *Journal of Research on Educational Effectiveness* **10** 297–325.
- EGGER, M., SMITH, G. D., SCHNEIDER, M. and MINDER, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315** 629–634.
- FELLER, A., GRINDAL, T., MIRATRIX, L. and PAGE, L. C. (2016a). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Ann. Appl. Stat.* **10** 1245–1285. [MR3553224](#)
- FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016b). Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models. Available at [arXiv:1602.06595](#).
- FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. [MR1891039](#)
- GELMAN, A., PARK, D. K., ANSOLABEHERE, S., PRICE, P. N. and MINNITE, L. C. (2001). Models, assumptions and model checking in ecological regressions. *J. Roy. Statist. Soc. Ser. A* **164** 101–118. [MR1819025](#)
- HULL, P. (2018). IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons. Working paper.
- JIANG, Z., DING, P. and GENG, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 829–848. [MR3534352](#)
- JO, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *J. Educ. Behav. Stat.* **27** 385–409.
- KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. [MR3494648](#)
- KLINE, P. and WALTERS, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *Q. J. Econ.* **131** 1795–1848.
- KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econom. Statist.* **33** 474–484. [MR3416595](#)
- MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.
- MEALLI, F., PACINI, B. and STANGHELLINI, E. (2016). Identification of principal causal effects using additional outcomes in concentration graphs. *J. Educ. Behav. Stat.* **41** 463–480.
- MIRATRIX, L., FUREY, J., FELLER, A., GRINDAL, T. and PAGE, L. C. (2018). Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness* **11** 133–162.
- PAGE, L. C., FELLER, A., GRINDAL, T., MIRATRIX, L. and SOMERS, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation* **36** 514–531.
- PECK, L. R. (2003). Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. *American Journal of Evaluation* **24** 157–187.

- RAUDENBUSH, S. W. and BLOOM, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation* **36** 475–499.
- RAUDENBUSH, S. W., REARDON, S. F. and NOMI, T. (2012). Statistical analysis for multisite trials using instrumental variables with random coefficients. *Journal of Research on Educational Effectiveness* **5** 303–332.
- RAUDENBUSH, S. and SCHWARTZ, D. (2017). Identification and estimation in multisite randomized trials with heterogeneous treatment effects. Submitted.
- REARDON, S. F. and RAUDENBUSH, S. W. (2013). Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociol. Methods Res.* **42** 143–163. [MR3190727](#)
- REARDON, S. F., UNLU, F., ZHU, P. and BLOOM, H. S. (2014). Bias and bias correction in multisite instrumental variables analysis of heterogeneous mediator effects. *J. Educ. Behav. Stat.* **39** 53–86.
- RUBIN, D. B. (1980). Comment on “Randomization analysis of experimental data: The Fisher randomization test”. *J. Amer. Statist. Assoc.* **75** 591–593.
- RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. [MR0899519](#)
- WANG, L., ZHOU, X.-H. and RICHARDSON, T. S. (2017). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika* **104** 597–612. [MR3694585](#)
- YUAN, L.-H. (2018). Regressions for estimating main and principal causal effects in multi-site randomized trials and small sample designs. Doctoral dissertation. Available at <https://dash.harvard.edu/handle/1/40050155>.
- YUAN, L.-H., FELLER, A. and MIRATRIX, L. (2019). Supplement to “Identifying and Estimating Principal Causal Effects in a Multi-site Trial of Early College High Schools.” DOI:[10.1214/18-AOAS1235SUPP](#).

IMPUTATION AND POST-SELECTION INFERENCE IN MODELS WITH MISSING DATA: AN APPLICATION TO COLORECTAL CANCER SURVEILLANCE GUIDELINES¹

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.

REFERENCES

- CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. [MR2243417](#)
- CHATTERJEE, A. and LAHIRI, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc. Amer. Math. Soc.* **138** 4497–4509. [MR2680074](#)
- CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. [MR2847974](#)
- CLAESKENS, G. (2016). Statistical model choice. *Ann. Rev. Stat. Appl.* **3** 233–256.
- CLAESKENS, G. and CONSENTINO, F. (2008). Variable selection with incomplete covariate data. *Biometrics* **64** 1062–1069. [MR2522253](#)
- CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. [MR2041482](#)
- CLAESKENS, G. and HJORT, N. L. (2008a). Minimizing average risk in regression models. *Econometric Theory* **24** 493–527. [MR2422864](#)
- CLAESKENS, G. and HJORT, N. L. (2008b). *Model Selection and Model Averaging*. *Cambridge Series in Statistical and Probabilistic Mathematics* **27**. Cambridge Univ. Press, Cambridge. [MR2431297](#)
- EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. [MR3265671](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- HEYMANS, M. W., VAN BUUREN, S., KNOL, D. L., VAN MECHELEN, W. and DE VET, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med. Res. Methodol.* **7**.
- HJORT, N. L. (2014). Comment [MR3265671]. *J. Amer. Statist. Assoc.* **109** 1017–1020. [MR3265676](#)
- HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. [MR2041481](#)
- HOSMER, D. W. and LEMESHOW, S. (1989). *Applied Logistic Regression*. Wiley-Interscience, New York.
- JONES, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J. Amer. Statist. Assoc.* **91** 222–230. [MR1394076](#)
- LACHENBRUCH, P. A. (2011). Variable selection when missing values are present: A case study. *Stat. Methods Med. Res.* **20** 429–444. [MR2829120](#)
- LIEBERMAN, D. A., REX, D. K., WINAWER, S. J., GIARDIELLO, F. M., JOHNSON, D. A. and LEVIN, T. R. (2012). Guidelines for colonoscopy surveillance after screening and polypectomy: A consensus update by the US multi-society task force on colorectal cancer. *Gastroenterology* **143** 844–857.
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. [MR1925014](#)
- LIU, L., MESSER, K., BARON, J. A., LIEBERMAN, D. A., JACOBS, E. T., CROSS, A. J., MURPHY, G., MARTINEZ, M. E. and GUPTA, S. (2016). A prognostic model for advanced colorectal neoplasia recurrence. *Cancer Causes Control* **27** 1175–1185. DOI:[10.1007/s10552-016-0795-5](#).
- LIU, L., QIU, Y., NATARAJAN, L. and MESSER, K. (2019). Supplement to “Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines.” DOI:[10.1214/19-AOAS1239SUPP](#).
- LONG, Q. and JOHNSON, B. A. (2015). Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics* **16** 596–610. [MR3365449](#)
- MARTINEZ, M. E., THOMPSON, P., MESSER, K. et al. (2012). One-year risk of advanced colorectal neoplasia: United States vs. United Kingdom risk-stratification guidelines. *Ann. Intern. Med.* **12** 856–864.

- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. [MR2758523](#)
- RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York. [MR0899519](#)
- SCHOMAKER, M. and HEUMANN, C. (2014). Model selection and model averaging after multiple imputation. *Comput. Statist. Data Anal.* **71** 758–770. [MR3132004](#)
- SCHOMAKER, M. and HEUMANN, C. (2018). Bootstrap inference when using multiple imputation. *Stat. Med.* **37** 2252–2266. [MR3810720](#)
- SIEGEL, R. L., MILLER, K. D. and JEMAL, A. (2015). Cancer statistics. *CA Cancer J. Clin.* **65** 5–29.
- TANNER, M. A. and WONG, W. H. (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics* **29** 23–32.
- TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York. [MR2233926](#)
- VAN BUUREN, S. and GROOTHUIS-ODDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#)
- WOOD, A. M., WHITE, I. R. and ROYSTON, P. (2008). How should variable selection be performed with multiply imputed data? *Stat. Med.* **27** 3227–3246. [MR2523914](#)

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

- BAUM, L. E. and EAGON, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73** 360–363. [MR0210217](#)
- BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37** 1554–1563. [MR0202264](#)
- BAUM, L. E. and SELL, G. R. (1968). Growth transformations for functions on manifolds. *Pacific J. Math.* **27** 211–227. [MR0234494](#)
- BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. [MR0287613](#)

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

- BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.
- CHING, W., FUNG, E. and NG, M. (2003). Higher-order hidden Markov models with applications to DNA sequences. In *Intelligent Data Engineering and Automated Learning. IDEAL 2003* (J. Liu, Y. Cheung and H. Yin, eds.). *Lecture Notes in Computer Science* **2690** 535–539. Springer, Berlin, Heidelberg.
- COLQUHOUN, D., HATTON, C. J. and HAWKES, A. G. (2003). The quality of maximum likelihood estimates of ion channel rate constants. *J. Physiol.* **547** 699–728.
- COLQUHOUN, D. and HAWKES, A. G. (1981). On the stochastic properties of single ion channels. *Proc. R. Soc. Lond., B Biol. Sci.* **211** 205–235.
- COLQUHOUN, D., HAWKES, A. G. and SRODZINSKI, K. (1996). Joint distributions of apparent open and shut times of single-ion channels and maximum likelihood fitting of mechanisms. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **354** 2555–2590.
- COX, S., ROSTEN, E., MONYPENNY, J., JOVANOVIĆ-TALISMAN, T., BURNETTE, D. T., LIPPINCOTT-SCHWARTZ, J., JONES, G. E. and HEINTZMANN, R. (2011). Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nat. Methods* **9** 195–200.
- DEMPSEY, G. T., VAUGHAN, J. C., CHEN, K. H., BATES, M. and ZHUANG, X. (2011). Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods* **8** 1027–1036.
- DU PREEZ, J. (1998). Efficient training of high-order hidden Markov models using first-order representations. *Comput. Speech Lang.* **12** 23–39.
- EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. [MR1270903](#)
- EPSTEIN, M., CALDERHEAD, B., GIROLAMI, M. A. and SIVILOTTI, L. (2016). Bayesian statistical inference in ion-channel models with exact missed event correction. *Biophys. J.* **111** 333–348.
- GREENFELD, M., PAVLICHIN, D. S., MABUCHI, H. and HERSCHLAG, D. (2015). Single molecule analysis research tool (SMART): An integrated approach for analysing single molecule data. *PLoS ONE* **7** e30024.
- HA, T. and TINNEFELD, P. (2012). Photophysics of fluorescent probes for single-molecule biophysics and super-resolution imaging. *Annu. Rev. Phys. Chem.* **63** 595–617.
- HAWKES, A. G., JALALI, A. and COLQUHOUN, D. (1990). The distributions of the apparent open times and shut times in a single channel record when brief events cannot be detected. *Philos. Trans. R. Soc. Lond. Ser. A* **332** 511–538. [MR1084721](#)
- HAWKES, A. G., JALALI, A. and COLQUHOUN, D. (1992). Asymptotic distributions of apparent open times and shut times in a single channel record allowing for the omission of brief events. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **337** 383–404.
- HEILEMANN, M., VAN DE LINDE, S., SCHÜTTPELZ, M., KASPER, R., SEEFELDT, B., MUKHERJEE, A., TINNEFELD, P. and SAUER, M. (2008). Subdiffraction—resolution fluorescence imaging with conventional fluorescent probes. *Angew. Chem. Int. Ed.* **47** 6172–6176.
- HESS, S. T., GIRIRAJAN, T. P. K. and MASON, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91** 4258–4272.
- HUANG, B., BATES, M. and ZHUANG, X. (2009). Super-resolution fluorescence microscopy. *Annu. Rev. Biochem.* **78** 993–1016.
- JACQUEZ, J. A. and GREIF, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Math. Biosci.* **77** 201–227.
- JUNGSMANN, R., STEINHÄUER, C., SCHEIBLE, M., KUZYK, A., TINNEFELD, P. and SIMMEL, F. C. (2010). Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. *Nano Lett.* **10** 4756–4761.

- LEE, L.-M. and LEE, J.-C. (2006). A study on high-order hidden Markov models and applications to speech recognition. In *Advances in Applied Artificial Intelligence. IEA/AIE 2006* (M. Ali and R. Dapoigny, eds.). *Lecture Notes in Computer Science* **4031** 682–690. Springer, Berlin, Heidelberg.
- LEE, S. H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441.
- LEHMANN, M., LICHTNER, G., KLENZ, H. and SCHMORANZER, J. (2016). Novel organic dyes for multicolor localization-based super-resolution microscopy. *J. Biophotonics* **9** 161–170.
- LEVINSON, S. E., RABINER, L. R. and SONDDHI, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Syst. Tech. J.* **62** 1035–1074. [MR0702893](#)
- LIN, Y., LONG, J. J., HUANG, F., DUIM, W. C., KIRSCHBAUM, S., ZHANG, Y., SCHROEDER, L. K., REBANE, A. A., VELASCO, M. G. M. et al. (2015). Quantifying and optimizing single-molecule switching nanoscopy at high speeds. *PLoS ONE* **10** e0128135.
- LITTLE, M. P., HEIDENREICH, W. F. and LI, G. (2010). Quantifying and optimizing single-molecule switching nanoscopy at high speeds. *PLoS ONE* **5** e8915.
- LIU, Y.-Y., LI, S., LI, F., SONG, L. and REHG, J. (2015). Efficient learning of continuous-time hidden Markov models for disease progression. In *NIPS Proceedings* 3600–3608.
- MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series. Monographs on Statistics and Applied Probability* **70**. CRC Press, London. [MR1692202](#)
- MUKAMEL, E., BABCOCK, H. and ZHUANG, X. (2012). Statistical deconvolution for superresolution fluorescence microscopy. *Biophys. J.* **102** 2391–2400.
- NIEUWENHUIZEN, R. P. J., BATES, M., SZYMBORSKA, A., LIDKE, K. A., RIEGER, B. and STALLINGA, S. (2015). Quantitative localization microscopy: Effects of photophysics and labeling stoichiometry. *PLoS ONE* **10** e0127989.
- OBER, R. J., RAM, S. and WARD, E. S. (2004). Localization accuracy in single-molecule microscopy. *Biophys. J.* **87** 1185–1200.
- OBER, R., TAHMASBI, A., RAM, S., LIN, Z. and WARD, E. (2015). Quantitative aspects of single-molecule microscopy: Information-theoretic analysis of single-molecule data. *IEEE Signal Process. Mag.* **32** 58–69.
- PATEL, L., GUSTAFSSON, N., LIN, Y., OBER, R., HENRIQUES, R. and COHEN, E. (2019). Supplement to “A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores.” DOI:[10.1214/19-AOAS1240SUPPA](#), DOI:[10.1214/19-AOAS1240SUPPB](#).
- QIN, F., AUERBACH, A. and SACHS, F. (1996). Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophys. J.* **70** 264–280.
- QIN, F. and LI, L. (2004). Model-based fitting of single-channel dwell-time distributions. *Biophys. J.* **87** 1657–1671.
- RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.
- RAM, S., WARD, E. S. and OBER, R. J. (2013). A stochastic analysis of distance estimation approaches in single molecule microscopy: Quantifying the resolution limits of photon-limited imaging systems. *Multidimens. Syst. Signal Process.* **24** 503–542. [MR3041619](#)
- RIEF, M., ROCK, R. S., MEHTA, A. D., MOOSEKER, M. S., CHENEY, R. E. and SPUDICH, J. A. (2000). Myosin-V stepping kinetics: A molecular model for processivity. *Proc. Natl. Acad. Sci. USA* **97** 9482–9486.
- RIEGER, B. and STALLINGA, S. (2014). The lateral and axial localization uncertainty in super-resolution light microscopy. *ChemPhysChem* **15** 664–670.
- ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2014). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** 110–118.

- ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591. [MR0436944](#)
- RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* 793–795.
- SAGE, D., KIRSHNER, H., PENG, T., STUURMAN, N., MIN, J., MANLEY, S. and USHER, M. (2015). Quantitative evaluation of software packages for single-molecule localization microscopy. *Nat. Methods* **12** 717–724.
- SHARONOV, A. and HOCHSTRASSER, R. M. (2006). Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proc. Natl. Acad. Sci. USA* **103** 18911–18916.
- THOMPSON, R. E., LARSON, D. R. and WEBB, W. W. (2002). Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82** 2775–2783.
- VAN DE LINDE, S. and SAUER, M. (2014). How to switch a fluorophore: From undesired blinking to controlled photoswitching. *Chem. Soc. Rev.* **43** 1076–1087.
- VAN DE LINDE, S., WOLTER, S., HEILEMANN, M. and SAUER, M. (2010). The effect of photo-switching kinetics and labeling densities on super-resolution fluorescence imaging. *J. Biotechnol.* **149** 260–266.
- VOGELSANG, J., STEINHÄUER, C., FORTHMANN, C., STEIN, I. H., PERSON-SKEGRO, B., CORDES, T. and TINNEFELD, P. (2010). Make them blink: Probes for super-resolution microscopy. *ChemPhysChem* **11** 2475–2490.

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

BY JENG-MIN CHIOU^{*,1}, 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 subsequences 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

- ASTON, J. A. D. and KIRCH, C. (2012). Detecting and estimating changes in dependent functional data. *J. Multivariate Anal.* **109** 204–220. [MR2922864](#)
- AUE, A., GABRYS, R., HORVÁTH, L. and KOKOSZKA, P. (2009). Estimation of a change-point in the mean function of functional data. *J. Multivariate Anal.* **100** 2254–2269. [MR2560367](#)
- BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. [MR2488343](#)
- BERKES, I., GABRYS, R., HORVÁTH, L. and KOKOSZKA, P. (2009). Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 927–946. [MR2750251](#)
- BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications. Lecture Notes in Statistics* **149**. Springer, New York. [MR1783138](#)
- BRAUN, J. V., BRAUN, R. K. and MÜLLER, H.-G. (2000). Multiple changepoint fitting via quasilielihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. [MR1782480](#)
- CHIOU, J.-M., CHEN, Y.-T. and HSING, T. (2019). Supplement to “Identifying multiple changes for a functional data sequence with application to freeway traffic segmentation.” DOI:[10.1214/19-AOAS1242SUPP](#).

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

- CHIOU, J.-M., CHEN, Y.-T. and YANG, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statist. Sinica* **24** 1571–1596. [MR3308652](#)
- CIUPERCA, G. (2011). A general criterion to determine the number of change-points. *Statist. Probab. Lett.* **81** 1267–1275. [MR2803773](#)
- FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics*. Springer, New York. [MR2229687](#)
- FREMDT, S., STEINEBACH, J. G., HORVÁTH, L. and KOKOSZKA, P. (2013). Testing the equality of covariance operators in functional samples. *Scand. J. Stat.* **40** 138–152. [MR3024036](#)
- FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. [MR3210728](#)
- FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. [MR3269979](#)
- GROMENKO, O., KOKOSZKA, P. and REIMHERR, M. (2017). Detection of change in the spatiotemporal mean function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 29–50. [MR3597963](#)
- HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc.* **113** 649–659. [MR3832216](#)
- HARCHAOU, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. [MR2796565](#)
- HÖRMANN, S. and KOKOSZKA, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845–1884. [MR2662361](#)
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications. Springer Series in Statistics*. Springer, New York. [MR2920735](#)
- HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR3379106](#)
- HU, S. R. and WANG, C. M. (2008). Vehicle detector deployment strategies for the estimation of network origin-destination demands using link traffic counts. *IEEE Transactions on Intelligent Transportation Systems* **9** 283–300.
- KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. [MR3036418](#)
- MAY, A. D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, NJ.
- NEWBY, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica* **59** 1161–1167. [MR1113551](#)
- NIU, Y. S., HAO, N. and ZHANG, H. (2016). Multiple change-point detection: A selective overview. *Statist. Sci.* **31** 611–623. [MR3598742](#)
- NIU, Y. S. and ZHANG, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6** 1306–1326. [MR3012531](#)
- OLSHEN, A. B., VENKATRAMAN, E., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.
- PAN, J. and CHEN, J. (2006). Application of modified information criterion to multiple change point problems. *J. Multivariate Anal.* **97** 2221–2241. [MR2301636](#)
- PANARETOS, V. M., KRAUS, D. and MADDOCKS, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Amer. Statist. Assoc.* **105** 670–682. [MR2724851](#)
- PÖTSCHER, B. M. and PRUCHA, I. R. (1994). Generic uniform convergence and equicontinuity concepts for random functions: An exploration of the basic structure. *J. Econometrics* **60** 23–63. [MR1247816](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](#)
- TRANSPORTATION RESEARCH BOARD (2010). *HCM 2010: Highway Capacity Manual*. TRB, Washington, DC.

- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer, New York. [MR1385671](#)
- WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3** 257–295.
- YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. [MR0919373](#)
- ZHANG, J.-T. (2014). *Analysis of Variance for Functional Data. Monographs on Statistics and Applied Probability* **127**. CRC Press, Boca Raton, FL. [MR3185072](#)

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

- AIZER, A. and DOYLE, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Q. J. Econ.* **130** 759–803.
- BHULLER, M., DAHL, G., LOKEN, K. and MOGSTAD, M. (2016). Incarceration, recidivism and employment. NBER 22648.
- CATTANEO, M., FRANDSEN, B. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *J. Causal Inference* **3** 1–24.
- CAUGHEY, D. and SEKHON, J. (2011). Elections and the regression discontinuity design: Lessons from close US House races. *Polit. Anal.* **19** 385–408.
- CHEN, H. and SMALL, D. (2016). New multivariate tests for assessing covariate balance in matched observational studies. Preprint. Available at [arXiv:1609.03686](https://arxiv.org/abs/1609.03686).
- DOBBIE, W., GOLDIN, J. and YANG, C. (2016). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. Working Paper No. 22511, National Bureau of Economic Research.
- DOYLE, J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *Am. Econ. Rev.* **97** 1583–1610.
- DOYLE, J. (2008). Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care. *J. Polit. Econ.* **116** 1201–1205.
- EGGERS, A. and HAINMUELLER, J. (2009). MPs for sale? Returns to office in postwar British politics. *Am. Polit. Sci. Rev.* **103** 513–533.

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

- GAGNON-BARTSCH, J. and SHEM-TOV, Y. (2019). Supplement to “The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies.” DOI:[10.1214/19-AOAS1241SUPPA](https://doi.org/10.1214/19-AOAS1241SUPPA), DOI:[10.1214/19-AOAS1241SUPPB](https://doi.org/10.1214/19-AOAS1241SUPPB).
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GREEN, D. P. and WINIK, D. (2010). Using random judge assignment to estimate the effect of incarceration and probation on recidivism among drug offenders. *Criminology* **48** 357–387.
- GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. [MR2913716](#)
- HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.* **23** 219–236. [MR2516821](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2722294](#)
- HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. [MR3068450](#)
- HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *Amer. Statist.* **64** 299–309. [MR2758561](#)
- HELLER, R., JENSEN, S. T., ROSENBAUM, P. R. and SMALL, D. S. (2010). Sensitivity analysis for the cross-match test, with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1005–1013. [MR2752596](#)
- LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. and GNEITING, T. (2017). Forecaster’s dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32** 106–127. [MR3634309](#)
- LUDWIG, J., MULLAINATHAN, S. and SPIESS, J. (2017). Machine learning tests for effects on multiple outcomes. Preprint. Available at [arXiv:1707.01473](https://arxiv.org/abs/1707.01473).
- LYALL, J. (2009). Does indiscriminate violence incite insurgent attacks?: Evidence from Chechnya. *J. Confl. Resolut.* **53** 331–362.
- ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17** 141–159. [MR0981441](#)
- ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. [MR2168202](#)
- ROSENBAUM, P. R. (2010). *Design of Observational Studies. Springer Series in Statistics*. Springer, New York. [MR2561612](#)
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. [MR0742974](#)
- ROUSE, C. (1995). Democratization or diversion? The effect of community colleges on educational attainment. *J. Bus. Econom. Statist.* **300** 217–224.
- STEVENSON, M. (2016). Distortion of justice: How the inability to pay bail affects case outcomes. Working paper. Available at <http://www.econ.pitt.edu/sites/default/files/Stevenson.jmp2016.pdf>.
- SZÉKELY, G. J. and RIZZO, M. L. (2009a). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. [MR2752127](#)
- SZÉKELY, G. J. and RIZZO, M. L. (2009b). Rejoinder: Brownian distance covariance. *Ann. Appl. Stat.* **3** 1303–1308. [MR2752135](#)
- TANG, Z.-Z., CHEN, G. and ALEKSEYENKO, A. V. (2016). PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* **32** 2618–2625.
- TASKINEN, S., OJA, H. and RANDELS, R. H. (2005). Multivariate nonparametric tests of independence. *J. Amer. Statist. Assoc.* **100** 916–925. [MR2201019](#)
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327](#)

SPATIO-TEMPORAL SHORT-TERM WIND FORECAST: A CALIBRATED REGIME-SWITCHING METHOD¹

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

- ADAMS, R. P. and MACKAY, D. J. (2007). Bayesian online changepoint detection. Preprint. Available at [arXiv:0710.3742](https://arxiv.org/abs/0710.3742).
- AGOSTINELLI, C. and LUND, U. (2013). R package “circular”: Circular statistics (version 0.4-7). <https://r-forge.r-project.org/projects/circular>.
- AILLIOT, P., MONBET, V. and PREVOSTO, M. (2006). An autoregressive model with time-varying coefficients for wind fields. *Environmetrics* **17** 107–117. [MR2225530](#)
- AILLIOT, P., BESSAC, J., MONBET, V. and PÈNE, F. (2015). Non-homogeneous hidden Markov-switching models for wind time series. *J. Statist. Plann. Inference* **160** 75–88. [MR3315635](#)
- ALEXIADIS, M., DOKOPOULOS, P., SAHSAMANOGLU, H. and MANOUSARIDIS, I. (1998). Short-term forecasting of wind speed and related electrical power. *Solar Energy* **63** 61–68.
- BESSA, R. J., MIRANDA, V., BOTTERUD, A., WANG, J. and CONSTANTINESCU, E. M. (2012). Time adaptive conditional kernel density estimation for wind power forecasting. *IEEE Trans. Sustain. Energy* **3** 660–669.
- BESSAC, J., AILLIOT, P., CATTIAUX, J. and MONBET, V. (2016). Comparison of hidden and observed regime-switching autoregressive models for (u, v)-components of wind fields in the north-east Atlantic. *Advances in Statistical Climatology, Meteorology and Oceanography* **2** 1–16.
- BROWELL, J., DREW, D. R. and PHILIPPOPOULOS, K. (2018). Improved very short-term spatio-temporal wind forecasting using atmospheric regimes. *Wind Energy* **21** 968–979.

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

- BROWN, B. G., KATZ, R. W. and MURPHY, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology* **23** 1184–1195.
- BYON, E. and DING, Y. (2010). Season-dependent condition-based maintenance for a wind turbine using a partially observed Markov decision process. *IEEE Trans. Power Syst.* **25** 1823–1834.
- BYON, E., NTAIMO, L. and DING, Y. (2010). Optimal maintenance strategies for wind turbine systems under stochastic weather conditions. *IEEE Trans. Reliab.* **59** 393–404.
- DE LUNA, X. and GENTON, M. G. (2005). Predictive spatio-temporal models for spatially sparse environmental data. *Statist. Sinica* **15** 547–568. [MR2190219](#)
- DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.
- DOE (2015). Wind Vision: A New Era for Wind Power in the United States. U.S. Department of Energy, Washington DC.
- DOWELL, J. and PINSON, P. (2016). Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Trans. Smart Grid* **7** 763–770.
- ERDEM, E. and SHI, J. (2011). ARMA based approaches for forecasting the tuple of wind speed and direction. *Applied Energy* **88** 1405–1414.
- EZZAT, A. A., JUN, M. and DING, Y. (2018). Spatio-temporal asymmetry of local wind fields and its impact on short-term wind forecasting. *IEEE Trans. Sustain. Energy* **9** 1437–1447.
- FUENTES, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics* **12** 469–483.
- GIEBEL, G., BROWNSWORD, R., KARINIOTAKIS, G., DENHARD, M. and DRAXL, C. (2011). The state-of-the-art in short-term prediction of wind power: A literature overview, 2nd ed. Technical report, ANEMOS.plus.
- GNEITING, T. (2002). Nonseparable, stationary covariance functions for space–time data. *J. Amer. Statist. Assoc.* **97** 590–600. [MR1941475](#)
- GNEITING, T. (2011a). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. [MR2847988](#)
- GNEITING, T. (2011b). Quantiles as optimal point forecasts. *Int. J. Forecast.* **27** 197–207.
- GNEITING, T., GENTON, M. G. and GUTTORP, P. (2007). Geostatistical space–time models, stationarity, separability and full symmetry. In *Statistics of Spatio-Temporal Systems* (B. Finkenstaedt, L. Held and V. Isham, eds.). *Monographs in Statistics and Applied Probability* **4** 151–175 4. CRC Press, Boca Raton, FL.
- GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *The Annual Review of Statistics and Its Application* **1** 125–151.
- GNEITING, T., LARSON, K., WESTRICK, K., GENTON, M. G. and ALDRICH, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *J. Amer. Statist. Assoc.* **101** 968–979. [MR2324108](#)
- HE, M., YANG, L., ZHANG, J. and VITAL, V. (2014). A spatio-temporal analysis approach for short-term forecast of wind farm generation. *IEEE Trans. Power Syst.* **29** 1611–1622.
- HERING, A. S. and GENTON, M. G. (2010). Powering up with space–time wind forecasting. *J. Amer. Statist. Assoc.* **105** 92–104. [MR2757195](#)
- HERING, A. S., KAZOR, K. and KLEIBER, W. (2015). A Markov-switching vector autoregressive stochastic wind generator for multiple spatial and temporal scales. *Resources* **4** 70–92.
- HUANG, Z. and CHALABI, Z. S. (1995). Use of time-series analysis to model and forecast wind speed. *J. Wind Eng. Ind. Aerodyn.* **56** 311–322.
- HWANGBO, H., JOHNSON, A. L. and DING, Y. (2017). A production economics analysis for quantifying the efficiency of wind turbines. *Wind Energy* **20** 1501–1513.
- HWANGBO, H., JOHNSON, A. L. and DING, Y. (2018). Spline model for wake effect analysis: Characteristics of a single wake and its impacts on wind turbine power generation. *IIEE Trans.* **50** 112–125.

- IEC (2005). Wind Turbines-Part 12-1: Power Performance Measurements of Electricity Producing Wind Turbines. IEC-International Electrotechnical Commission, IEC-61400-12, Geneva, Switzerland.
- JAMMALAMADAKA, S. R. and SENGUPTA, A. (2001). *Topics in Circular Statistics. Series on Multivariate Analysis* **5**. World Scientific, River Edge, NJ. [MR1836122](#)
- JUN, M. and STEIN, M. L. (2007). An approach to producing space–time covariance functions on spheres. *Technometrics* **49** 468–479. [MR2394558](#)
- KAZOR, K. and HERING, A. S. (2015a). The role of regimes in short-term wind speed forecasting at multiple wind farms. *Stat* **4** 271–290. [MR3429323](#)
- KAZOR, K. and HERING, A. S. (2015b). Assessing the performance of model-based clustering methods in multivariate time series with application to identifying regional wind regimes. *J. Agric. Biol. Environ. Stat.* **20** 192–217. [MR3347625](#)
- KILLICK, R. and ECKLEY, I. (2014). changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **58** 1–19.
- KUSIAK, A. and LI, W. (2010). Estimation of wind speed: A data-driven approach. *Journal of Wind Engineering and Industrial Aerodynamics* **98** 559–567.
- LEE, G., DING, Y., GENTON, M. G. and XIE, L. (2015). Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Amer. Statist. Assoc.* **110** 56–67. [MR3338486](#)
- MOHANDES, M. A., HALAWANI, T. O., REHMAN, S. and HUSSAIN, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy* **29** 939–947.
- PINSON, P. (2013). Wind energy: Forecasting challenges for its operational management. *Statist. Sci.* **28** 564–585. [MR3161588](#)
- PINSON, P., CHEVALLIER, C. and KARINIOTAKIS, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Trans. Power Syst.* **22** 1148–1156.
- PINSON, P., CHRISTENSEN, L., MADSEN, H., SØRENSEN, P. E., DONOVAN, M. H. and JENSEN, L. E. (2008). Regime-switching modelling of the fluctuations of offshore wind generation. *J. Wind Eng. Ind. Aerodyn.* **96** 2327–2347.
- POURHABIB, A., HUANG, J. Z. and DING, Y. (2016). Short-term wind speed forecast using measurements from multiple turbines in a wind farm. *Technometrics* **58** 138–147. [MR3463164](#)
- REIKARD, G. (2008). Using temperature and state transitions to forecast wind speed. *Wind Energy* **11** 431–443.
- SAATÇI, Y., TURNER, R. and RASMUSSEN, C. E. (2010). Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning, (ICML-10)* 927–934.
- SANTOS, R. A. (2007). Damage mitigating control for wind turbines. Ph.D. thesis. Univ. Colorado Boulder, Boulder, CO.
- SCHLATHER, M. (2010). Some covariance models based on normal scale mixtures. *Bernoulli* **16** 780–797. [MR2730648](#)
- SIDERATOS, G. and HATZIARGYRIOU, N. D. (2012). Probabilistic wind power forecasting using radial basis function neural networks. *IEEE Trans. Power Syst.* **27** 1788–1796.
- STEIN, M. L. (2005). Space–time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. [MR2156840](#)
- XIE, L., GU, Y., ZHU, X. and GENTON, M. G. (2014). Short-term spatio-temporal wind power forecast in robust look-ahead power system dispatch. *IEEE Trans. Smart Grid* **5** 511–520.
- YOU, M., BYON, E., JIN, J. and LEE, G. (2017). When wind travels through turbines: A new statistical approach for characterizing heterogeneous wake effects in multi-turbine wind farms. *IIEE Trans.* **49** 84–95.
- ZHU, X., GENTON, M. G., GU, Y. and XIE, L. (2014). Space–time wind speed forecasting for improved power system dispatch. *TEST* **23** 1–25. [MR3179610](#)
- ZWIERKS, F. and VON STORCH, H. (1990). Regime-dependent autoregressive time series modeling of the southern oscillation. *J. Climate* **3** 1347–1363.

NETWORK MODELLING OF TOPOLOGICAL DOMAINS USING HI-C DATA

BY Y. X. RACHEL WANG^{*,1}, 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

- BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.
- CABREROS, I., ABBE, E. and TSIRIGOS, A. (2016). Detecting community structures in hi-c genomic data. In *Information Science and Systems (CISS), 2016 Annual Conference on* 584–589. IEEE Press, New York.
- DEKKER, J. (2008). Gene regulation in the third dimension. *Science* **319** 1793–1794.
- DIXON, J. R. et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485** 376–380.
- ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.
- FILIPPOVA, D., PATRO, R., DUGGAL, G. and KINGSFORD, C. (2014). Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9** 14.

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

- HOU, C., LI, L., QIN, Z. S. and CORCES, V. G. (2012). Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Molecular Cell* **48** 471–484.
- KELLIS, M. et al. (2014). Defining functional dna elements in the human genome. *Proc. Natl. Acad. Sci. USA* **111** 6131–6138.
- KNIGHT, P. A. and RUIZ, D. (2013). A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33** 1029–1047. [MR3081493](#)
- LE DILY, F. et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes & Development* **28** 2151–2162.
- LÉVY-LEDUC, C., DELATTRE, M., MARY-HUARD, T. and ROBIN, S. (2014). Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics* **30** i386–i392.
- LIEBERMAN-AIDEN, E. et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326** 289–293.
- LUPIÁÑEZ, D. G. et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161** 1012–1025.
- MALIK, L. I. and PATRO, R. (2015). Rich chromatin structure prediction from hi-c data. *bioRxiv*, page 032953.
- MEABURN, K. J., GUDLA, P. R., KHAN, S., LOCKETT, S. J. and MISTELI, T. (2009). Disease-specific gene repositioning in breast cancer. *J. Cell Biol.* **187** 801–812.
- NORA, E. P. et al. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* **485** 381–385.
- NORTON, H. K., EMERSON, D. J., HUANG, H., KIM, J., TITUS, K. R., GU, S., BASSETT, D. S. and PHILLIPS-CREMINS, J. E. (2018). Detecting hierarchical genome folding with network modularity. *Nat. Methods* **15** 119–122.
- RAO, S. S. et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159** 1665–1680.
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. [MR2893856](#)
- SANBORN, A. L., RAO, S. S. P., HUANG, S.-C., DURAND, N. C., HUNTLEY, M. H., JEWETT, A. I., BOCHKOV, I. D., CHINNAPPAN, D., CUTKOSKY, A. et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112** E6456–E6465.
- SAURIA, M. E., PHILLIPS-CREMINS, J. E., CORCES, V. G. and TAYLOR, J. (2014). Hifive: A normalization approach for higher-resolution hic and 5c chromosome conformation data analysis. Available at <https://www.biorxiv.org/content/10.1101/009951v1.full>.
- SEXTON, T. et al. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148** 458–472.
- SMITH, E. M., LAJOIE, B. R., JAIN, G. and DEKKER, J. (2016). Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the CFTR locus. *Am. J. Hum. Genet.* **98** 185–201.
- WANG, Y. X. R., SARKAR, P., URSU, O., KUNDAJE, A. and BICKEL, P. J. (2019). Supplement to “Network modelling of topological domains using Hi-C data.” DOI:10.1214/19-AOAS1244SUPP.
- WEINREB, C. and RAPHAEL, B. J. (2016). Identification of hierarchical chromatin domains. *Bioinformatics* **32** 1601–1609.
- YAN, K.-K., LOU, S. and GERSTEIN, M. (2017). MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput. Biol.* **13** e1005647.

FAST DYNAMIC NONPARAMETRIC DISTRIBUTION TRACKING IN ELECTRON MICROSCOPIC DATA¹

BY YANJUN QIAN*, JIANHUA Z. HUANG[†], CHIWOO 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, crystalizing 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

- ALDOUS, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation and coagulation): A review of the mean-field theory for probabilists. *Bernoulli* **5** 3–48. [MR1673235](#)
- ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254. [MR0028556](#)
- BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. [MR0381130](#)
- BOAL, A. K., ILHAN, F., DEROUCHÉY, J. E., THURN-ALBRECHT, T., RUSSELL, T. P. and ROTELLO, V. M. (2000). Self-assembly of nanoparticles into structured spherical and network aggregates. *Nature* **404** 746–748.
- BROWN, L., CAI, T., ZHANG, R., ZHAO, L. and ZHOU, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields* **146** 401–433. [MR2574733](#)
- DE JONG, P. and SHEPHARD, N. (1995). The simulation smoother for time series models. *Biometrika* **82** 339–350. [MR1354233](#)

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

- DOUCET, A., GORDON, N. J. and KRISHNAMURTHY, V. (2001). Particle filters for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.* **49** 613–624.
- DURBIN, J. and KOOPMAN, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84** 669–684. [MR1603944](#)
- EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with *B*-splines and penalties. *Statist. Sci.* **11** 89–121. [MR1435485](#)
- GRZELCZAK, M., VERMANT, J., FURST, E. M. and LIZ-MARZÁN, L. M. (2010). Directed self-assembly of nanoparticles. *ACS Nano* **4** 3591–3605.
- KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45.
- LI, M., SCHNABLEGGER, H. and MANN, S. (1999). Coupled synthesis and self-assembly of nanoparticles to give structures with controlled organization. *Nature* **402** 393–395.
- LIFSHITZ, I. and SLYOZOV, V. (1961). The kinetics of precipitation from supersaturated solid solutions. *J. Phys. Chem. Solids* **19** 35–50.
- LJUNG, L. (1979). Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Trans. Automat. Control* **24** 36–50. [MR0519391](#)
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. [MR0733519](#)
- MA, J., KOCKELMAN, K. M. and DAMIEN, P. (2008). A multivariate Poisson-lognormal regression model for prediction of crash counts by severity using Bayesian methods. *Accident Anal. Prev.* **40** 964–975.
- MAHALANOBIS, P. C. (1993). On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2** 49–55.
- MENA, R. H. and RUGGIERO, M. (2016). Dynamic density estimation with diffusive Dirichlet mixtures. *Bernoulli* **22** 901–926. [MR3449803](#)
- MUNEESAWANG, P. and SIRISATHITKUL, C. (2015). Size measurement of nanoparticle assembly using multilevel segmented TEM images. *J. Nanomater.* **16** 58–63.
- PARK, C. (2014). Estimating multiple pathways of object growth using nonlongitudinal image data. *Technometrics* **56** 186–199. [MR3207846](#)
- PARK, C., HUANG, J., HUITINK, D., KUNDU, S., MALLICK, B., LIANG, H. and DING, Y. (2012). A multi-stage, semi-automated procedure for analyzing the morphology of nanoparticles. *IIE Trans.* **44** 507–522.
- PARK, C., HUANG, J., JI, J. and DING, Y. (2013). Segmentation, inference and classification of partially overlapping nanoparticles. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 669–681.
- PARK, C., WOHL, T. J., EVANS, J. E. and BROWNING, N. D. (2015). Minimum cost multi-way data association for optimizing multitarget tracking of interacting objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 611–624.
- QIAN, Y., HUANG, J. Z. and DING, Y. (2017). Identifying multi-stage nanocrystal growth using in situ TEM video data. *IIE Trans.* **49** 532–543.
- QIAN, Y., HUANG, J. Z., LI, X. and DING, Y. (2016). Robust nanoparticles detection from noisy background by fusing complementary image information. *IEEE Trans. Image Process.* **25** 5713–5726. [MR3562077](#)
- QIAN, Y., HUANG, J. Z., PARK, C. and DING, Y. (2019). Supplement to “Fast dynamic nonparametric distribution tracking in electron microscopic data.” DOI:[10.1214/19-AOAS1245SUPPA](#), DOI:[10.1214/19-AOAS1245SUPPB](#).
- RODRIGUEZ, A. and TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Anal.* **3** 339–365. [MR2407430](#)
- SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690. [MR1125725](#)
- SIMONOFF, J. S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Ann. Statist.* **11** 208–218. [MR0684878](#)

- SPIEGELHALTER, D., THOMAS, A., BEST, N. and GILKS, W. (1996). BUGS 0.5: Bayesian inference using Gibbs sampling manual (versio ii). In *MRC Biostatistics Unit* 1–59. Institute of Public Health, Cambridge, UK.
- WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. [MR1045442](#)
- WOEHL, T. J., PARK, C., EVANS, J. E., ARSLAN, I., RISTENPART, W. D. and BROWNING, N. D. (2013). Direct observation of aggregative nanoparticle growth: Kinetic modeling of the size distribution and growth rate. *Nano Lett.* **14** 373–378.
- ZHANG, C., CHEN, N. and LI, Z. (2017). State space modeling of autocorrelated multivariate Poisson counts. *IJSE Trans.* **49** 518–531.
- ZHENG, H., SMITH, R. K., JUN, Y. W., KISIELOWSKI, C., DAHMEN, U. and ALIVISATOS, A. P. (2009). Observation of single colloidal platinum nanocrystal growth trajectories. *Science* **324** 1309–1312.

DISTRIBUTIONAL REGRESSION FORESTS FOR PROBABILISTIC PRECIPITATION FORECASTING IN COMPLEX TERRAIN

BY LISA SCHLOSSER*, TORSTEN HOTHORN^{†,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

- ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. [MR3909963](#)
- BARAN, S. and NEMODA, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* **27** 280–292. [MR3521202](#)
- BAUER, P., THORPE, A. and BRUNET, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** (7567) 47–55.
- BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. [MR3493512](#)

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

- BMLFUW (2016). Bundesministerium für Land und Forstwirtschaft, Umwelt und Wasserwirtschaft (BMLFUW), Abteilung IV/4—Wasserhaushalt. Available at <http://ehyd.gv.at/>. Accessed: 2016–02–29.
- BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. [MR0192611](#)
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. [MR0726392](#)
- DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.
- GEBETSBERGER, M., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Fine-tuning non-homogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Weather Rev.* **145** 4693–4708.
- GLAHN, H. R. and LOWRY, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11** 1203–1211.
- GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. [MR2325275](#)
- GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. [MR2345548](#)
- GNEITING, T., RAFTERY, A. E., WESTVELD III, A. H. and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.
- GOICOA, T., ADIN, A., UGARTE, M. D. and HODGES, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stoch. Environ. Res. Risk Assess.* **32** 749–770.
- HAMILL, T. M., BATES, G. T., WHITAKER, J. S., MURRAY, D. R., FIORINO, M., GALARNEAU JR., T. J., ZHU, Y. and LAPENTA, W. (2013). NOAA’s second-generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94** 1553–1565.
- HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. [MR0858512](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York. [MR1851606](#)
- HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.
- HOFNER, B., MAYR, A. and SCHMID, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *J. Stat. Softw.* **74** (1) 1–31.
- HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. [MR2291267](#)
- HOTHORN, T. and ZEILEIS, A. (2017). Transformation forests. Available at [arXiv:1701.02110](https://arxiv.org/abs/1701.02110).
- HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91.
- HOTHORN, T., HORNIK, K., VAN DE WIEL, M. A. and ZEILEIS, A. (2006). A Lego system for conditional inference. *Amer. Statist.* **60** 257–263. [MR2246759](#)
- HUTCHINSON, M. F. (1998). Interpolation of rainfall data with thin plate smoothing splines—Part II: Analysis of topographic dependence. *Journal of Geographic Information and Decision Analysis* **2** 152–167.
- KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.* **9** 1024–1052. [MR3371346](#)

- LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *J. Amer. Statist. Assoc.* **101** 578–590. [MR2256176](#)
- LONG, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks, CA.
- MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. [MR2274394](#)
- MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2016). Heteroscedastic censored and truncated regression with crch. *The R Journal* **8** (1) 173–181.
- MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017). Non-homogeneous boosting for predictor selection in ensemble post-processing. *Mon. Weather Rev.* **145** 137–147.
- NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A* **135** 370–384.
- RASP, S. and LERCH, S. (2018). Neural networks for post-processing ensemble weather forecasts. *Mon. Weather Rev.* **146** 3885–3900.
- RIGBY, R. A. and STASINOPOULOS, D. M. (2005a). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. [MR2137253](#)
- ROBINSON, N., REGETZ, J. and GURALNICK, R. P. (2014). EarthEnv-DEM90: A nearly-global, void-free, multi-scale smoothed, 90m digital elevation model from fused ASTER and SRTM data. *ISPRS J. Photogramm. Remote Sens.* **87** 57–67.
- SCHEUERER, M. and HAMILL, T. M. (2015). Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Weather Rev.* **143** 4578–4596.
- SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019a). Different response distributions. Supplement A to “Distributional regression forests for probabilistic precipitation forecasting in complex terrain.” DOI:[10.1214/19-AOAS1247SUPPA](#).
- SCHLOSSER, L., HOTHORN, T., STAUFFER, R. and ZEILEIS, A. (2019b). Stationwise evaluation. Supplement B to “Distributional regression forests for probabilistic precipitation forecasting in complex terrain.” DOI:[10.1214/19-AOAS1247SUPPB](#).
- STASINOPOULOS, D. M. and RIGBY, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23** (7) 1–46.
- STAUFFER, R., UMLAUF, N., MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2017a). Ensemble post-processing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Weather Rev.* **45** 955–969.
- STAUFFER, R., MAYR, G. J., MESSNER, J. W., UMLAUF, N. and ZEILEIS, A. (2017b). Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.* **37** 3264–3275.
- STIDD, C. K. (1973). Estimating the precipitation climate. *Water Resour. Res.* **9** 1235–1241.
- STRASSER, H. and WEBER, CH. (1999). The asymptotic theory of permutation statistics. *Math. Methods Statist.* **8** 220–250. Johann Pfanzagl—On the occasion of his 70th birthday. [MR1722622](#)
- UGARTE, M. D., ADIN, A. and GOICOA, T. (2017b). One-dimensional, two-dimensional, and three dimensional B-splines to specify space-time interactions in Bayesian disease mapping: Model fitting and model identifiability. *Spat. Stat.* **22** 451–468. [MR3732861](#)
- WOOD, S. N., SCHEIPL, F. and FARAWAY, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat. Comput.* **23** 341–360. [MR3041440](#)
- ZEILEIS, A. and HORNIK, K. (2007). Generalized M -fluctuation tests for parameter instability. *Stat. Neerl.* **61** 488–508. [MR2351461](#)
- ZEILEIS, A., HOTHORN, T. and HORNIK, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Statist.* **17** 492–514. [MR2439970](#)

MODELING SEASONALITY AND SERIAL DEPENDENCE OF ELECTRICITY PRICE CURVES WITH WARPING FUNCTIONAL AUTOREGRESSIVE DYNAMICS¹

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

- AMJADY, N., DARAEPOUR, A. and KEYNIA, F. (2010). Day-ahead electricity price forecasting by modified relief algorithm and hybrid neural network. *IET Gener. Transm. Distrib.* **4** 432–444.
- AMJADY, N. and KEYNIA, F. (2009). Day-ahead price forecasting of electricity markets by mutual information technique and cascaded neuro-evolutionary algorithm. *IEEE Trans. Power Syst.* **24** 306–318.
- ANTONIADIS, A. and SAPATINAS, T. (2003). Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J. Multivariate Anal.* **87** 133–158. [MR2007265](#)
- BELLO, A., RESESES, J., MUÑOZ, A. and DELGADILLO, A. (2016). Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques. *Int. J. Forecast.* **32** 966–980.
- BESSE, P., CARDOT, H. and STEPHENSON, D. (2000). Autoregressive forecasting of some climatic variations. *Scand. J. Stat.* **27** 673–687.
- BHATIA, R. and HOLBROOK, J. (2006). Riemannian geometry and matrix geometric means. *Linear Algebra Appl.* **413** 594–618. [MR2198952](#)
- BOOGERT, A. and DUPONT, D. (2008). When supply meets demand: The case of hourly spot electricity prices. *IEEE Trans. Power Syst.* **23** 389–398.
- BOSQ, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 509–529. Springer, Berlin.
- BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications. Lecture Notes in Statistics* **149**. Springer, New York. [MR1783138](#)

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

- ČENCŮV, N. N. (1982). *Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs* **53**. Amer. Math. Soc., Providence, RI. [MR0645898](#)
- CHEN, Y. and LI, B. (2017). An adaptive functional autoregressive forecast model to predict electricity price curves. *J. Bus. Econom. Statist.* **35** 371–388. [MR3663780](#)
- CONTRERAS, J., ESPINOLA, R., NOGALES, F. J. and CONEJO, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.* **18** 1014–1020.
- DIDERICKSEN, D., KOKOSZKA, P. and ZHANG, X. (2012). Empirical properties of forecasts with the functional autoregressive model. *Comput. Statist.* **27** 285–298. [MR2923229](#)
- DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.
- EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. [MR0428531](#)
- FENG, Q., HANNIG, J. and MARRON, J. (2016). A note on automatic data transformation. *Stat* **5** 82–87.
- GAILLARD, P., GOUDE, Y. and NEDELLEC, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *Int. J. Forecast.* **32** 1038–1050.
- GARCIA-MARTOS, C., RODRIGUEZ, J. and SÁNCHEZ, M. (2012). Forecasting electricity prices by extracting dynamic common factors: Application to the Iberian market. *IET Gener. Transm. Distrib.* **6** 11–20.
- GASSER, T. and KNEIP, A. (1995). Searching for structure in curve samples. *J. Amer. Statist. Assoc.* **90** 1179–1188.
- GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414. [MR0653512](#)
- GRENNANDER, U. (1981). *Abstract Inference*. Wiley, New York. [MR0599175](#)
- HERNÁEZ, E. C., GIL, J. B., LEÓN, J. D. L. F., SAN ROQUE, A., RODRIGUEZ, M., GONZÁLEZ, J., GONZÁLEZ, A. and CALMARZA, A. (2006). Competitors’ response representation for market simulation in the Spanish daily market. In *Modelling Prices in Competitive Electricity Markets* 21–68. Wiley, New York.
- HOWISON, S. and COULON, M. (2009). Stochastic behaviour of the electricity bid stack: From fundamental drivers to power prices. *J. Energy Mark.* **2** 29–69.
- JANCZURA, J., TRÜCK, S., WERON, R. and WOLFF, R. C. (2013). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Econ.* **38** 96–110.
- JOSKOW, P. L. (2001). California’s electricity crisis. *Oxford Rev. Econ. Policy* **17** 365–388.
- JUBAN, R., OHLSSON, H., MAASOUMY, M., POIRIER, L. and KOLTER, J. Z. (2016). A multiple quantile regression approach to the wind, solar, and price tracks of GEFCom2014. *Int. J. Forecast.* **32** 1094–1102.
- KANAMURA, T. and ŌHASHI, K. (2007). A structural model for electricity prices with spikes: Measurement of spike risk and optimal policies for hydropower plant operation. *Energy Econ.* **29** 1010–1032.
- KANAMURA, T. and ŌHASHI, K. (2008). On transition probabilities of regime switching in electricity prices. *Energy Econ.* **30** 1158–1172.
- KASS, R. E. and VOS, P. W. (1997). *Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics: Probability and Statistics* **908**. Wiley, New York.
- KNEIP, A. and GASSER, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20** 1266–1305. [MR1186250](#)
- KOKOSZKA, P. and ZHANG, X. (2010). Improved estimation of the kernel of the functional autoregressive process. Technical report, Utah State Univ., Logan, UT.
- KOOPMAN, S. J., OOMS, M. and CARNERO, M. A. (2007). Periodic seasonal Reg-ARFIMA-GARCH models for daily electricity spot prices. *J. Amer. Statist. Assoc.* **102** 16–27. [MR2345529](#)

- LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.* **7** 1562–1592. [MR3127959](#)
- LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. [MR2172368](#)
- MACIEJOWSKA, K. and NOWOTARSKI, J. (2016). A hybrid model for GEFCom2014 probabilistic electricity price forecasting. *Int. J. Forecast.* **32** 1051–1056.
- MACIEJOWSKA, K., NOWOTARSKI, J. and WERON, R. (2016). Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *Int. J. Forecast.* **32** 957–965.
- MACIEJOWSKA, K. and WERON, R. (2016). Short-and mid-term forecasting of baseload electricity prices in the UK: The impact of intra-day price relationships and market fundamentals. *IEEE Trans. Power Syst.* **31** 994–1005.
- MARRON, J. S., RAMSAY, J. O., SANGALLI, L. M. and SRIVASTAVA, A. (2014). Statistics of time warping and phase variations. *Electron. J. Stat.* **8** 1697–1702. [MR3273584](#)
- MARRON, J. S., RAMSAY, J. O., SANGALLI, L. M. and SRIVASTAVA, A. (2015). Functional data analysis of amplitude and phase variation. *Statist. Sci.* **30** 468–484. [MR3432837](#)
- MOURID, T. and BENSMAN, N. (2006). Sieves estimator of the operator of a functional autoregressive process. *Statist. Probab. Lett.* **76** 93–108. [MR2213247](#)
- NOGALES, F. J., CONTRERAS, J., CONEJO, A. J. and ESPÍNOLA, R. (2002). Forecasting next-day electricity prices by time series models. *IEEE Trans. Power Syst.* **17** 342–348.
- NOWOTARSKI, J. and WERON, R. (2015). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Comput. Statist.* **30** 791–803. [MR3404360](#)
- PAPE, C., HAGEMANN, S. and WEBER, C. (2016). Are fundamentals enough? Explaining price variations in the German day-ahead and intraday power market. *Energy Econ.* **54** 376–387.
- RADHAKRISHNA RAO, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** 81–91. [MR0015748](#)
- RAMSAY, J. O. and LI, X. (1998). Curve registration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 351–363. [MR1616045](#)
- SAKOE, H. and CHIBA, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26** 43–49.
- SHAHIDEPOUR, M., YAMIN, H. and LI, Z. (2002). *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management*. Wiley, New York.
- SKANTZE, P. L. and ILIC, M. D. (2012). *Valuation, Hedging and Speculation in Competitive Electricity Markets: A Fundamental Approach*. Springer, Berlin.
- SRIVASTAVA, A., WU, W., KURTEK, S., KLASSEN, E. and MARRON, J. S. (2011). Registration of functional data using Fisher–Rao metric. arXiv preprint. Available at [arXiv:1103.3817](#).
- TANG, R. and MÜLLER, H.-G. (2008). Pairwise curve synchronization for functional data. *Biometrika* **95** 875–889. [MR2461217](#)
- UNIEJEWSKI, B., WERON, R. and ZIEL, F. (2018). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Trans. Power Syst.* **33** 2219–2229.
- VEHVILÄINEN, I. and PYYKKÖNEN, T. (2005). Stochastic factor model for electricity spot price—the case of the Nordic market. *Energy Econ.* **27** 351–367.
- WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.
- WANG, K. and GASSER, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.* **25** 1251–1276. [MR1447750](#)
- WERON, R. (2007). *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. Wiley.
- WERON, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *Int. J. Forecast.* **30** 1030–1081.
- WERON, R. and MISIOREK, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *Int. J. Forecast.* **24** 744–763.

- ZIEL, F., STEINERT, R. and HUSMANN, S. (2015). Efficient modeling and forecasting of electricity spot prices. *Energy Econ.* **47** 98–111.
- ZIEL, F. and WERON, R. (2018). Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate models. *Energy Econ.* **70** 396–420.

RCRNORM: AN INTEGRATED SYSTEM OF RANDOM-COEFFICIENT HIERARCHICAL REGRESSION MODELS FOR NORMALIZING NANOSTRING NCOUNTER DATA¹

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

- ABDUEVA, D., WING, M., SCHAUB, B., TRICHE, T. and DAVICIONI, E. (2010). Quantitative expression profiling in formalin-fixed paraffin-embedded samples by affymetrix microarrays. *J. Mol. Diagnostics* **12** 409–417.
- APRIL, C., KLOTZLE, B., ROYCE, T., WICKHAM-GARCIA, E., BOYANIWSKY, T., IZZO, J., COX, D., JONES, W., RUBIO, R. et al. (2009). Whole-genome gene expression profiling of formalin-fixed, paraffin-embedded tissue samples. *PLoS ONE* **4** e8162.
- CHEN, H.-Y., YU, S.-L., CHEN, C.-H., CHANG, G.-C., CHEN, C.-Y., YUAN, A., CHENG, C.-L., WANG, C.-H., TERNG, H.-J. et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.* **356** 11–20.

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

- EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. [MR0615434](#)
- EISENBERG, E. and LEVANON, E. Y. (2013). Human housekeeping genes, revisited. *Trends Genet.* **29** 569–574.
- FRIEDMAN, M. M. (1997). Clinical laboratory improvement amendment. *Home Healthc. Now* **15** 393–395.
- GEISS, G. K., BUMGARNER, R. E., BIRDITT, B., DAHL, T., DOWIDAR, N., DUNAWAY, D. L., FELL, H. P., FERREE, S., GEORGE, R. D. et al. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.* **26** 317–325.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](#)
- GUBERN, C., HURTADO, O., RODRÍGUEZ, R., MORALES, J. R., ROMERA, V. G., MORO, M. A., LIZASOAIN, I., SERENA, J. and MALLOLAS, J. (2009). Validation of housekeeping genes for quantitative real-time PCR in in-vivo and in-vitro models of cerebral ischaemia. *BMC Mol. Biol.* **10** 57.
- IDDAWELA, M., RUEDA, O. M., KLARQVIST, M., GRAF, S., EARL, H. M. and CALDAS, C. (2016). Reliable gene expression profiling of formalin-fixed paraffin-embedded breast cancer tissue (FFPE) using cDNA-mediated annealing, extension, selection, and ligation whole-genome (DASL WG) assay. *BMC Med. Genomics* **9** 54.
- IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.
- JIA, G., WANG, X., LI, Q., LU, W., TANG, X., WISTUBA, I. and XIE, Y. (2019). Supplement to “RCRnorm: An integrated system of random-coefficient hierarchical regression models for normalizing NanoString nCounter data.” DOI:[10.1214/19-AOAS1249SUPP](#).
- KULKARNI, M. M. (2011). Digital multiplexed gene expression analysis using the NanoString nCounter system. *Curr. Protoc. Mol. Biol.* **94** 25B.10.1-25B.10.17.
- LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.
- LI, C. and WONG, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98** 31–36.
- LÜDER RIPOLI, F., MOHR, A., CONRADINE HAMMER, S., WILLENBROCK, S., HEWICKER-TRAUTWEIN, M., HENNECKE, S., MURUA ESCOBAR, H. and NOLTE, I. (2016). A comparison of fresh frozen vs. formalin-fixed, paraffin-embedded specimens of canine mammary tumors via branched-DNA assay. *Int. J. Mol. Sci.* **17** 724.
- LUDWIG, J. A. and WEINSTEIN, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* **5** 845–856.
- MASUDA, N., OHNISHI, T., KAWAMOTO, S., MONDEN, M. and OKUBO, K. (1999). Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Res.* **27** 4436–4443.
- NANOSTRING TECHNOLOGIES, INC. (2017). Gene Expression Data Analysis Guidelines—NanoString. MAN-C0011-04.
- OMOLO, B., YANG, M., LO, F. Y., SCHELL, M. J., AUSTIN, S., HOWARD, K., MADAN, A. and YEATMAN, T. J. (2016). Adaptation of a RAS pathway activation signature from FF to FFPE tissues in colorectal cancer. *BMC Med. Genomics* **9** 65.
- PALUCH, B. E., GLENN, S. T., CONROY, J. M., PAPANICOLAOU-SENGOS, A., BSHARA, W., OMILIAN, A. R., BRESE, E., NESLINE, M., BURGHER, B. et al. (2017). Robust detection of immune transcripts in FFPE samples using targeted RNA sequencing. *Oncotarget* **8** 3197–3205.

- PERLMUTTER, M. A., BEST, C. J., GILLESPIE, J. W., GATHRIGHT, Y., GONZÁLEZ, S., VELASCO, A., LINEHAN, W. M., EMMERT-BUCK, M. R. and CHUAQUI, R. F. (2004). Comparison of snap freezing versus ethanol fixation for gene expression profiling of tissue specimens. *J. Mol. Diagnostics* **6** 371–377.
- REIS, P. P., WALDRON, L., GOSWAMI, R. S., XU, W., XUAN, Y., PEREZ-ORDONEZ, B., GULLANE, P., IRISH, J., JURISICA, I. et al. (2011). mRNA transcript quantification in archival samples using multiplexed, color-coded probes. *BMC Biotechnol.* **11** 46.
- RISSO, D., NGAI, J., SPEED, T. P. and DUDOIT, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32** 896.
- ROSENFELD, N., AHARONOV, R., MEIRI, E., ROSENWALD, S., SPECTOR, Y., ZEPENIUK, M., BENJAMIN, H., SHABES, N., TABAK, S. et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.* **26** 462–469.
- SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika* **52** 591–611. [MR0205384](#)
- SOLASSOL, J., RAMOS, J., CRAPEZ, E., SAIFI, M., MANGÉ, A., VIANÈS, E., LAMY, P.-J., COSTES, V. and MAUDELONDE, T. (2011). Kras mutation detection in paired frozen and formalin-fixed paraffin-embedded (FFPE) colorectal cancer tissues. *Int. J. Mol. Sci.* **12** 3191–3204.
- STEFAN, U., MICHAEL, B. and WERNER, S. (2010). Effects of three different preservation methods on the mechanical properties of human and bovine cortical bone. *Bone* **47** 1048–1053.
- TANG, H., XIAO, G., BEHRENS, C., SCHILLER, J., ALLEN, J., CHOW, C.-W., SURAKAR, M., CORVALAN, A., MAO, J. et al. (2013). A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin. Cancer Res.* **19** 1577–1586.
- THOMPSON, D., BOTROS, I., ROUNSEVILLE, M., LIU, Q., WANG, E., HARRISON, H. and ROCHE, P. (2014). Automated high fidelity RNA expression profiling using nuclease protection coupled with next generation sequencing. In *Merck Technology Symposium, Long Branch, NJ* **6**.
- VALLEJOS, C. A., RISSO, D., SCIALDONE, A., DUDOIT, S. and MARIONI, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14** 565–571.
- VON AHLFEN, S., MISSEL, A., BENDRAT, K. and SCHLUMBERGER, M. (2007). Determinants of RNA quality from FFPE samples. *PLoS ONE* **2** e1261.
- WAGGOTT, D., CHU, K., YIN, S., WOUTERS, B. G., LIU, F.-F. and BOUTROS, P. C. (2012). NanoStringNorm: An extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinformatics* **28** 1546–1548.
- WANG, H., HORBINSKI, C., WU, H., LIU, Y., SHENG, S., LIU, J., WEISS, H., STROMBERG, A. J. and WANG, C. (2016). NanoStringDiff: A novel statistical method for differential expression analysis based on NanoString nCounter data. *Nucleic Acids Res.* **44** e151–e151.
- XIE, Y., XIAO, G., COOMBES, K. R., BEHRENS, C., SOLIS, L. M., RASO, G., GIRARD, L., ERICKSON, H. S., ROTH, J. et al. (2011). Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin. Cancer Res.* **17** 5705–5714.
- XIE, Y., LU, W., WANG, S., TANG, X., TANG, H., ZHOU, Y., MORAN, C., BEHRENS, C., ROTH, J. et al. (2017). Validation of the 12-gene predictive signature for adjuvant chemotherapy response in lung cancer. *J. Thorac. Oncol.* **12** S1544.
- ZIOBER, A. F., PATEL, K. R., ALAWI, F., GIMOTTY, P., WEBER, R. S., FELDMAN, M. M., CHALIAN, A. A., WEINSTEIN, G. S., HUNT, J. et al. (2006). Identification of a gene signature for rapid screening of oral squamous cell carcinoma. *Clin. Cancer Res.* **12** 5960–5971.

NETWORK CLASSIFICATION WITH APPLICATIONS TO BRAIN CONNECTOMICS¹

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

- AINE, C. J., BOCKHOLT, H. J., BUSTILLO, J. R., CAÑIVE, J. M., CAPRIHAN, A., GASPAROVIC, C., HANLON, F. M., HOUCK, J. M., JUNG, R. E. et al. (2017). Multimodal neuroimaging in schizophrenia: Description and dissemination. *Neuroinformatics* **15** 343–364.
- ARROYO RELIÓN, J. D., KESSLER, D., LEVINA, E. and TAYLOR, S. F. (2019). Supplement to “Network classification with applications to brain connectomics.” DOI:[10.1214/19-AOAS1252SUPPA](https://doi.org/10.1214/19-AOAS1252SUPPA), DOI:[10.1214/19-AOAS1252SUPPB](https://doi.org/10.1214/19-AOAS1252SUPPB).
- BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. [MR2417268](#)
- BACH, F., JENATTON, R., MAIRAL, J. and OBOZINSKI, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.* **27** 450–468. [MR3025128](#)
- BECK, A. and TEOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. [MR2486527](#)

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

- BECKER, N., WERFT, W., TOEDT, G., LICHTER, P. and BENNER, A. (2009). penalizedSVM: A R-package for feature selection SVM classification. *Bioinformatics* **25** 1711–1712.
- BENGIO, Y. and MONPERRUS, M. (2005). Non-local manifold tangent learning. In *Advances in Neural Information Processing Systems* 129–136.
- BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.
- BORGWARDT, K. M., ONG, C. S., SCHÖNAUER, S., VISHWANATHAN, S., SMOLA, A. J. and KRIEGEL, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics* **21** i47–i56.
- BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.
- BROYD, S. J., DEMANUELE, C., DEBENER, S., HELPS, S. K., JAMES, C. J. and SONUGA-BARKE, E. J. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neurosci. Biobehav. Rev.* **33** 279–296.
- BULLMORE, E. T. and BASSETT, D. S. (2011). Brain graphs: Graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* **7** 113–140.
- BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198.
- BUNNEY, W. E. and BUNNEY, B. G. (2000). Evidence for a compromised dorsolateral prefrontal cortical parallel circuit in schizophrenia. *Brains Res. Rev.* **31** 138–146.
- BUTTON, K. S., IOANNIDIS, J. P., MOKRYSZ, C., NOSEK, B. A., FLINT, J., ROBINSON, E. S. and MUNAFÒ, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev., Neurosci.* **14** 365–376.
- CHEN, X., LIN, Q., KIM, S., CARBONELL, J. G. and XING, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression. *Ann. Appl. Stat.* **6** 719–752. [MR2976489](#)
- CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.
- CRADDOCK, R. C., HOLTZHEIMER, P. E., HU, X. P. and MAYBERG, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.* **62** 1619–1628.
- DESHPANDE, M., KURAMOCHI, M., WALE, N. and KARYPIS, G. (2005). Frequent substructure-based approaches for classifying chemical compounds. *IEEE Trans. Knowl. Data Eng.* **17** 1036–1050.
- DONG, D., WANG, Y., CHANG, X., LUO, C. and YAO, D. (2017). Dysfunction of large-scale brain networks in schizophrenia: A meta-analysis of resting-state functional connectivity. *Schizophr. Bull.* **44** 168–181.
- DUCHI, J. and SINGER, Y. (2009). Efficient online and batch learning using forward backward splitting. *J. Mach. Learn. Res.* **10** 2899–2934. [MR2579916](#)
- DURANTE, D. and DUNSON, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Anal.* **13** 29–58. [MR3737942](#)
- FEI, H. and HUAN, J. (2010). Boosting with structure information in the functional space: An application to graph classification. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 643–652. ACM, New York.
- FINN, E. S., SHEN, X., SCHEINOST, D., ROSENBERG, M. D., HUANG, J., CHUN, M. M., PAPADEMETRIS, X. and CONSTABLE, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18** 1664–1671.
- FORNITO, A., ZALESKY, A., PANTELIS, C. and BULLMORE, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *NeuroImage* **62** 2296–2314.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010a). A note on the group lasso and a sparse group lasso. Available at [arXiv:1001.0736](#).
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

- GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. [MR3687603](#)
- GÄRTNER, T., FLACH, P. and WROBEL, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines* 129–143. Springer, Berlin.
- GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. [MR3693544](#)
- GONZALEZ, J., HOLDER, L. B. and COOK, D. J. (2000). Graph based concept learning. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press 1072. MIT Press, Cambridge, MA.
- GOTTS, S., SAAD, Z., JO, H. J., WALLACE, G., COX, R. and MARTIN, A. (2013). The perils of global signal regression for group comparisons: A case study of autism spectrum disorders. *Front. Human Neurosci.* **7** 356.
- GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **72** 304–321.
- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability **143**. CRC Press, Boca Raton, FL. [MR3616141](#)
- HELMA, C., KING, R. D., KRAMER, S. and SRINIVASAN, A. (2001). The predictive toxicology challenge 2000–2001. *Bioinformatics* **17** 107–108.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#)
- HU, Y. and ALLEN, G. I. (2015). Local-aggregate modeling for big data via distributed optimization: Applications to neuroimaging. *Biometrics* **71** 905–917. [MR3436716](#)
- INOKUCHI, A., WASHIO, T. and MOTODA, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *Principles of Data Mining and Knowledge Discovery* 13–23. Springer, Berlin.
- JACOB, L., OBOZINSKI, G. and VERT, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning* 433–440. ACM, New York.
- KASHIMA, H., TSUDA, K. and INOKUCHI, A. (2003). Marginalized kernels between labeled graphs. In *International Conference of Machine Learning* **3** 321–328.
- KETKAR, N. S., HOLDER, L. B. and COOK, D. J. (2009). Empirical comparison of graph classification algorithms. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on* 259–266. IEEE, New York.
- KIAR, G., BRIDGEFORD, E., RONCAL, W. G., CHANDRASHEKHAR, V., MHEMBERE, D., RYMAN, S., ZUO, X.-N., MARGUILES, D. S., CRADDOCK, R. C. et al. (2018). A high-throughput pipeline identifies robust connectomes but troublesome variability. *BioRxiv* 188706.
- KUDO, T., MAEDA, E. and MATSUMOTO, Y. (2004). An application of boosting to graph classification. In *Advances in Neural Information Processing Systems* 729–736.
- LE, C. M., LEVINA, E. and VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Structures Algorithms* **51** 538–561. [MR3689343](#)
- LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electron. J. Stat.* **9** 608–642. [MR3331852](#)
- LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. [MR3485948](#)
- LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. [MR2530545](#)
- LIU, Y., LIANG, M., ZHOU, Y., HE, Y., HAO, Y., SONG, M., YU, C., LIU, H., LIU, Z. et al. (2008). Disrupted small-world networks in schizophrenia. *Brain* **131** 945–961.

- LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. [MR3210970](#)
- MEINSHAUSEN, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393. [MR2409990](#)
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#)
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. [MR2758523](#)
- MENON, V. (2011). Large-scale brain networks and psychopathology: A unifying triple network model. *Trends Cogn. Sci.* **15** 483–506.
- NARAYAN, M., ALLEN, G. I. and TOMSON, S. (2015). Two sample inference for populations of graphical models with applications to functional connectivity. Available at [arXiv:1502.03853](#).
- ONGÜR, D., LUNDY, M., GREENHOUSE, I., SHINN, A. K., MENON, V., COHEN, B. M. and RENSHAW, P. F. (2010). Default mode network abnormalities in bipolar disorder and schizophrenia. *Psychiatry Res.* **183** 59–68.
- PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.
- PEETERS, S. C., VAN DE VEN, V., GRONENSCILD, E. H. M., PATEL, A. X., HABETS, P., GOEBEL, R., VAN OS, J. and MARCELIS, M. (2015). Default mode network connectivity as a function of familial and environmental risk for psychotic disorder. *PLoS ONE* **10** e0120030.
- POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.
- PRASAD, G., JOSHI, S. H., NIR, T. M., TOGA, A. W., THOMPSON, P. M. and ADNI (2015). Brain connectivity and novel network measures for Alzheimer’s disease classification. *Neurobiol. Aging* **36** 121–131.
- RICHIARDI, J., ERYILMAZ, H., SCHWARTZ, S., VUILLEUMIER, P. and VAN DE VILLE, D. (2011). Decoding brain states from fMRI connectivity graphs. *NeuroImage* **56** 616–626.
- SCHEINBERG, K., GOLDFARB, D. and BAI, X. (2014). Fast first-order methods for composite convex optimization with backtracking. *Found. Comput. Math.* **14** 389–417. [MR3201951](#)
- SCOTT, J. G., KELLY, R. C., SMITH, M. A., ZHOU, P. and KASS, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.* **110** 459–471. [MR3367240](#)
- SHAH, R. D. and SAMWORTH, R. J. (2013). Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 55–80. [MR3008271](#)
- SMITH, S. M., VIDAURRE, D., BECKMANN, C. F., GLASSER, M. F., JENKINSON, M., MILLER, K. L., NICHOLS, T. E., ROBINSON, E. C., SALIMI-KHORSHIDI, G. et al. (2013). Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17** 666–682.
- SRINIVASAN, A., MUGGLETON, S. H., STERNBERG, M. J. and KING, R. D. (1996). Theories for mutagenicity: A study in first-order and feature-based induction. *Artificial Intelligence* **85** 277–299.
- SRIPADA, C., ANGSTADT, M., KESSLER, D., PHAN, K. L., LIBERZON, I., EVANS, G. W., WELSH, R. C., KIM, P. and SWAIN, J. E. (2014a). Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks. *NeuroImage* **89** 110–121.
- SRIPADA, C., KESSLER, D., FANG, Y., WELSH, R. C., PREM KUMAR, K. and ANGSTADT, M. (2014b). Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder. *Hum. Brain Mapp.* **35** 4693–4705.
- SUPEKAR, K., MENON, V., RUBIN, D., MUSEN, M. and GREICIUS, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer’s disease. *PLoS Comput. Biol.* **4** e1000100.

- TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017a). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. [MR3640191](#)
- TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2017b). A non-parametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23** 1599–1630. [MR3624872](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. [MR3224285](#)
- VAROQUAUX, G. and CRADDOCK, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage* **80** 405–415.
- VISHWANATHAN, S. V. N., SCHRAUDOLPH, N. N., KONDOR, R. and BORGWARDT, K. M. (2010). Graph kernels. *J. Mach. Learn. Res.* **11** 1201–1242. [MR2645450](#)
- VOGELSTEIN, J. T., RONCAL, W. G., VOGELSTEIN, R. J. and PRIEBE, C. E. (2013). Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1539–1551.
- WATANABE, T., KESSLER, D., SCOTT, C., ANGSTADT, M. and SRIPADA, C. (2014). Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. *NeuroImage* **96** 183–202.
- WHITFIELD-GABRIELI, S., THERMENOS, H. W., MILANOVIC, S., TSUANG, M. T., FARAONE, S. V., MCCARLEY, R. W., SHENTON, M. E., GREEN, A. I., NIETO-CASTANON, A. et al. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proc. Natl. Acad. Sci. USA* **106** 1279–1284.
- XIN, B., KAWAHARA, Y., WANG, Y. and GAO, W. (2014). Efficient generalized fused lasso and its application to the diagnosis of Alzheimer’s disease. In *Twenty-Eighth AAAI Conference on Artificial Intelligence* 2163–2169.
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. [MR2212574](#)
- YUAN, L., LIU, J. and YE, J. (2011). Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems* 352–360.
- ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207.
- ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. [MR3546450](#)
- ZHANG, J., CHENG, W., WANG, Z., ZHANG, Z., LU, W., LU, G. and FENG, J. (2012). Pattern classification of large-scale functional brain networks: Identification of informative neuroimaging markers for epilepsy. *PLoS ONE* **7** e36733.
- ZHANG, L., GUINDANI, M., VERSACE, F., ENGELMANN, J. M. and VANNUCCI, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *Ann. Appl. Stat.* **10** 638–666. [MR3528355](#)
- ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. [MR3174640](#)
- ZHU, J., ROSSET, S., TIBSHIRANI, R. and HASTIE, T. J. (2004). 1-norm support vector machines. In *Advances in Neural Information Processing Systems* 49–56.
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327](#)

SEQUENTIAL DECISION MODEL FOR INFERENCE AND PREDICTION ON NONUNIFORM HYPERGRAPHS WITH APPLICATION TO KNOT MATCHING FROM COMPUTATIONAL FORESTRY¹

BY SEONG-HWAN JUN*, SAMUEL W. K. WONG[†], JAMES V. ZIDEK* 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

- ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. [MR2758115](#)
- BERG-KIRKPATRICK, T., BOUCHARD-CÔTÉ, A., DENERO, J. and KLEIN, D. (2010). Painless unsupervised learning with features. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL10)* **8** 582–590.
- BONDY, J. A. and MURTY, U. S. R. (1976). *Graph Theory with Applications*. Elsevier, New York. [MR0411988](#)
- BOUCHARD-CÔTÉ, A. and JORDAN, M. I. (2010). Variational inference over combinatorial spaces. In *Advances in Neural Information Processing Systems 23 (NIPS)* **23** 280–288.
- BOUCHARD-CÔTÉ, A., SANKARARAMAN, S. and JORDAN, M. I. (2012). Phylogenetic inference via sequential Monte Carlo. *Syst. Biol.* **61** 579–593. DOI:[10.1093/sysbio/syr131](#).
- BRÄNNSTRÖM, M. (2009). The impact of a strength grading process on sawmill profitability and product quality. *BioResources* **4** 1430–1454.
- CAETANO, T. S., MCAULEY, J. J., CHENG, L., LE, Q. V. and SMOLA, A. J. (2009). Learning graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **31** 1048–1058.
- CAO, Z., QIN, T., LIU, T., TSAI, M. and LI, H. (2007). Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning* 129–136. ACM, New York.
- CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181. [MR3262549](#)

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

- CASTÉRA, P., FAYE, C. and EL OUADRANI, A. (1996). Prevision of the bending strength of timber with a multivariate statistical approach. In *Annales des Sciences Forestières* **53** 885–896. EDP Sciences.
- DAVAL, V., POT, G., BELKACEMI, M., MERIAUDEAU, F. and COLLET, R. (2015). Automatic measurement of wood fiber orientation and knot detection using an optical system based on heating conduction. *Opt. Express* **23** 33529–33539.
- DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. [MR2278333](#)
- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](#)
- DOUCET, A. and JOHANSEN, A. M. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *The Oxford Handbook of Nonlinear Filtering* 656–704. Oxford Univ. Press, Oxford. [MR2884612](#)
- GREEN, D. W., ROSS, R. J. and McDONALD, K. A. (1994). Production of hardwood machine stress rated lumber. In *Proceedings of 9th International Symposium on Nondestructive Testing of Wood* 141–150.
- GUINDOS, P. and GUAITA, M. (2013). A three-dimensional wood material model to simulate the behavior of wood with any type of knot at the macro-scale. *Wood Sci. Technol.* **47** 585–599.
- HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. [MR2086387](#)
- HIETANIEMI, R., HANNUKSELA, J. and SILVEÉN, O. (2011). Camera based lumber strength classification system. In *MVA2011 IAPR Conference on Machine Vision Applications* 251–254.
- HIETANIEMI, R., LÓPEZ, M. B., HANNUKSELA, J. and SILVÉN, O. (2014). A real-time imaging system for lumber strength prediction. *For. Prod. J.* **64** 126–133.
- HOLMES, I. and RUBIN, G. M. (2001). Pairwise RNA structure comparison with stochastic context-free grammars. In *Proceedings of the Pac. Symp. Biocomputing* 7 163–174.
- JUN, S.-H., WONG, S. W., ZIDEK, J. and BOUCHARD-COTE, A. (2017). Sequential graph matching with sequential Monte Carlo. In *Artificial Intelligence and Statistics* 1075–1084.
- KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. [MR0075510](#)
- LEVANDOWSKY, M. and WINTER, D. (1971). Distance between sets. *Nature* **234** 34–35.
- LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. [MR1038245](#)
- LU, B. and ROSENBAUM, P. R. (2004). Optimal pair matching with two control groups. *J. Comput. Graph. Statist.* **13** 422–434. [MR2063993](#)
- LUNTER, G., DRUMMOND, A. J., MIKLÓS, I. and HEIN, J. (2005). Statistical alignment: Recent progress, new applications, and challenges. In *Statistical Methods in Molecular Evolution. Stat. Biol. Health* 375–405. Springer, New York. [MR2161838](#)
- MØLLER, J., PETTITT, A. N., REEVES, R. and BERTHELSEN, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93** 451–458. [MR2278096](#)
- OLSSON, A., OSCARSSON, J., SERRANO, E., KÄLLSNER, B., JOHANSSON, M. and ENQUIST, B. (2013). Prediction of timber bending strength and in-member cross-sectional stiffness variation on the basis of local wood fibre orientation. *Eur. J. Wood Prod.* **71** 319–333.
- PAPADIMITRIOU, C. H. and STEIGLITZ, K. (1982). *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Englewood Cliffs, NJ. [MR0663728](#)
- PETTERSON, J., YU, J., MCAULEY, J. J. and CAETANO, T. S. (2009). Exponential family graph matching and ranking. In *Advances in Neural Information Processing Systems* 1455–1463.
- PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. [MR0391338](#)

- WANG, L., BOUCHARD-CÔTÉ, A. and DOUCET, A. (2015). Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. *J. Amer. Statist. Assoc.* **110** 1362–1374. [MR3449032](#)
- WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.
- WICKHAM, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
- WONG, S. W. K., LUM, C., WU, L. and ZIDEK, J. V. (2016). Quantifying uncertainty in lumber grading and strength prediction: A Bayesian approach. *Technometrics* **58** 236–243. [MR3488302](#)

A BAYESIAN MARK INTERACTION MODEL FOR ANALYSIS OF TUMOR PATHOLOGY IMAGES¹

BY QIWEI LI*, XINLEI WANG[†], FAMING LIANG[‡] AND GUANGHUA XIAO^{§,2}

*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

- AMIN, M. B., TAMBOLI, P., MERCHANT, S. H., ORDÓÑEZ, N. G., RO, J., AYALA, A. G. and RO, J. Y. (2002). Micropapillary component in lung adenocarcinoma: A distinctive histologic feature with possible prognostic significance. *Amer. J. Surg. Pathol.* **26** 358–364.
- AVALOS, G. and BUCCI, F. (2014). Exponential decay properties of a mathematical model for a certain fluid-structure interaction. In *New Prospects in Direct, Inverse and Control Problems for Evolution Equations. Springer INdAM Ser.* **10** 49–78. Springer, Cham. [MR3362986](#)
- BADDELEY, A. (2010). Multivariate and marked point processes. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 371–402. CRC Press, Boca Raton, FL. [MR2730956](#)
- BADDELEY, A. and TURNER, R. (2006). Modelling spatial point patterns in *R*. In *Case Studies in Spatial Point Process Modeling. Lect. Notes Stat.* **185** 23–74. Springer, New York. [MR2232122](#)
- BARLETTA, J. A., YEAP, B. Y. and CHIRIEAC, L. R. (2010). Prognostic significance of grading in lung adenocarcinoma. *Cancer* **116** 659–669.

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

- BECK, A. H., SANGOI, A. R., LEUNG, S., MARINELLI, R. J., NIELSEN, T. O., VAN DE VIJVER, M. J., WEST, R. B., VAN DE RIJN, M. and KOLLER, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Translational Med.* **3** 108–113.
- BESAG, J. E. (1977). Comment on “Modelling spatial patterns.” *J. Roy. Statist. Soc. Ser. B* **39** 193–195.
- BOGNAR, M. A. (2008). Bayesian modeling of continuously marked spatial point patterns. *Comput. Statist.* **23** 361–379. [MR2425167](#)
- BORCZUK, A. C., QIAN, F., KAZEROS, A., ELEAZAR, J., ASSAAD, A., SONETT, J. R., GINSBURG, M., GORENSTEIN, L. and POWELL, C. A. (2009). Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. *Amer. J. Surg. Pathol.* **33** 462.
- BRAMBILLA, E., LE TEUFF, G., MARGUET, S., LANTUEJOUL, S., DUNANT, A., GRAZIANO, S., PIRKER, R., DOUILLARD, J.-Y., LE CHEVALIER, T., FILIPITS, M. et al. (2016). Prognostic effect of tumor lymphocytic infiltration in resectable non-small-cell lung cancer. *J. Clin. Oncol.* **34** 1223–1230.
- CARPENTER, A. E., JONES, T. R., LAMPRECHT, M. R., CLARKE, C., KANG, I. H., FRIMAN, O., GUERTIN, D. A., CHANG, J. H., LINDQUIST, R. A., MOFFAT, J. et al. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7** R100.
- CHULAEVSKY, V. (2014). Exponential decay of eigenfunctions in a continuous multi-particle Anderson model with sub-exponentially decaying interaction. Available at [arXiv:1408.4646](#).
- COX, T. F. (1979). A method for mapping the dense and sparse regions of a forest stand. *Appl. Statist.* **28** 14–19.
- COX, D. R. (1992). Regression models and life-tables In *Breakthroughs in Statistics, Vol. II. Methodology and Distribution* (S. Kotz and N. L. Johnson, eds.) 527–541. Springer, Berlin.
- COX, T. F. and LEWIS, T. (1976). A conditioned distance ratio method for analyzing spatial patterns. *Biometrika* **63** 483–491. [MR0445713](#)
- DALE, M. R. (2000). *Spatial Pattern Analysis in Plant Ecology*. Cambridge Univ. Press, Cambridge.
- DIGGLE, P. J. (1986). Displaced amacrine cells in the retina of a rabbit: Analysis of a bivariate spatial point pattern. *J. Neuroscience Methods* **18** 115–125.
- DIGGLE, P. J. and COX, T. (1981). On sparse sampling methods and tests of independence for multivariate spatial point patterns. *Bulletin Internat. Statist. Inst.* **49** 213–229.
- DIGGLE, P. J., EGLEN, S. J. and TROY, J. B. (2006). Modelling the bivariate spatial distribution of amacrine cells. In *Case Studies in Spatial Point Process Modeling. Lect. Notes Stat.* **185** 215–233. Springer, New York. [MR2232131](#)
- DIGGLE, P. J. and MILNE, R. K. (1983). Bivariate Cox processes: Some models for bivariate spatial point patterns. *J. Roy. Statist. Soc. Ser. B* **45** 11–21. [MR0701070](#)
- FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. [MR1951635](#)
- GELFAND, A. E., DIGGLE, P., GUTTORP, P. and FUENTES, M. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.
- GELMAN, A., RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](#)
- GILLIES, R. J., VERDUZCO, D. and GATENBY, R. A. (2012). Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nat. Rev. Cancer* **12** 487–493.
- GLEASON, D. F. and MELLINGER, G. T. (2002). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urology* **167** 953–958.
- GRABARNIK, P., MYLLYMÄKI, M. and STOYAN, D. (2011). Correct testing of mark independence for marked point patterns. *Ecol. Model.* **222** 3888–3894.

- HAMMERSLEY, J. M. and CLIFFORD, P. (1971). Markov fields on finite graphs and lattices.
- HANAHAN, D. and WEINBERG, R. A. (2011). Hallmarks of cancer: The next generation. *Cell* **144** 646–674.
- HUGHES, A. (1981a). Cat retina and the sampling theorem: The relation of transient and sustained brisk-unit cut-off frequency to α and β -mode cell density. *Experimental Brain Res.* **42** 196–202.
- HUGHES, A. (1981b). Population magnitudes and distribution of the major modal classes of cat retinal ganglion cell as estimated from HRP filling and a systematic survey of the soma diameter spectra for classical neurones. *J. Comparative Neurology* **197** 303–339.
- HUH, J. W., LEE, J. H. and KIM, H. R. (2012). Prognostic significance of tumor-infiltrating lymphocytes for patients with colorectal cancer. *Archives of Surgery* **147** 366–372.
- HUI, E. E. and BHATIA, S. N. (2007). Micromechanical control of cell–cell interactions. *Proc. Natn. Acad. Sci.* **104** 5722–5726.
- ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice.* Wiley, Chichester. [MR2384630](#)
- JUNTILLA, M. R. and DE SAUVAGE, F. J. (2013). Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501** 346–354.
- KAMENSKY, L., JONES, T. R., FRASER, A., BRAY, M.-A., LOGAN, D. J., MADDEN, K. L., LJOSA, V., RUEDEN, C., ELICEIRI, K. W. and CARPENTER, A. E. (2011). Improved structure, function and compatibility for CellProfiler: Modular high-throughput image analysis software. *Bioinformatics* **27** 1179–1180.
- KASHIMA, Y. (2010). Exponential decay of correlation functions in many-electron systems. *J. Math. Phys.* **51** 063521, 40. [MR2676498](#)
- LI, Q., YI, F., WANG, T., XIAO, G. and LIANG, F. (2017). Lung cancer pathological image analysis using a hidden Potts model. *Cancer Informatics* **16** 1176935117711910.
- LI, Q., WANG, X., LIANG, F. and XIAO, G. (2019). Supplement to “A Bayesian mark interaction model for analysis of tumor pathology images.” DOI:[10.1214/19-AOAS1254SUPPA](#), DOI:[10.1214/19-AOAS1254SUPPB](#).
- LIANG, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *J. Stat. Comput. Simul.* **80** 1007–1022. [MR2742519](#)
- LIANG, F., JIN, I. H., SONG, Q. and LIU, J. S. (2016). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *J. Amer. Statist. Assoc.* **111** 377–393. [MR3494666](#)
- LOTWICK, H. W. and SILVERMAN, B. W. (1982). Methods for analysing spatial processes of several types of points. *J. Roy. Statist. Soc. Ser. B* **44** 406–413. [MR0693241](#)
- LUO, X., ZANG, X., YANG, L., HUANG, J., LIANG, F., CANALES, J. R., WISTUBA, I. I., GAZDAR, A., XIE, Y. and XIAO, G. (2016). Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J. Thorac. Oncol.* DOI:[10.1016/j.jtho.2016.10.017](#).
- MANTOVANI, A., SOZZANI, S., LOCATI, M., ALLAVENA, P. and SICA, A. (2002). Macrophage polarization: Tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. *Trends in Immunology* **23** 549–555.
- MATTFELDT, T., ECKEL, S., FLEISCHER, F. and SCHMIDT, V. (2009). Statistical analysis of labelling patterns of mammary carcinoma cell nuclei on histological sections. *J. Microsc.* **235** 106–118. [MR2731112](#)
- MERLO, L. M., PEPPER, J. W., REID, B. J. and MALEY, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* **6** 924–935.
- MØLLER, J., PETTITT, A. N., REEVES, R. and BERTHELTSEN, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93** 451–458. [MR2278096](#)
- MURRAY, I., GHAHRAMANI, Z. and MACKAY, D. (2012). MCMC for doubly-intractable distributions. Available at [arXiv:1206.6848](#).

- OGATA, Y. and TANEMURA, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics* **41** 421–433.
- ORIMO, A., GUPTA, P. B., SGROI, D. C., ARENZANA-SEISDEDOS, F., DELAUNAY, T., NAEEM, R., CAREY, V. J., RICHARDSON, A. L. and WEINBERG, R. A. (2005). Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* **121** 335–348.
- PEICHL, L. and WÄSSLE, H. (1981). Morphological identification of on-and off-centre brisk transient (Y) cells in the cat retina. *Proc. Roy. Soc. London Ser. B, Biolog. Sci.* **212** 139–153.
- PENROSE, O. and LEBOWITZ, J. L. (1974). On the exponential decay of correlation functions. *Comm. Math. Phys.* **39** 165–184. [MR0432092](#)
- POLYAK, K., HAVIV, I. and CAMPBELL, I. G. (2009). Co-evolution of tumor cells and their microenvironment. *Trends Genet.* **25** 30–38.
- REN, Q., BANERJEE, S., FINLEY, A. O. and HODGES, J. S. (2011). Variational Bayesian methods for spatial data analysis. *Comput. Statist. Data Anal.* **55** 3197–3217. [MR2825404](#)
- RINCÓN, J., GANAHL, M. and VIDAL, G. (2015). Lieb–Liniger model with exponentially decaying interactions: A continuous matrix product state study. *Phys. Rev. B* **92** 115107.
- RIPLEY, B. D. (1977). Modelling spatial patterns. *J. Roy. Statist. Soc. Ser. B* **39** 172–212. [MR0488279](#)
- ROCKHILL, R. L., EULER, T. and MASLAND, R. H. (2000). Spatial order within but not between types of retinal neurons. *Proc. Natn. Acad. Sci.* **97** 2303–2307.
- SABO, E., BECK, A. H., MONTGOMERY, E. A., BHATTACHARYA, B., MEITNER, P., WANG, J. Y. and RESNICK, M. B. (2006). Computerized morphometry as an aid in determining the grade of dysplasia and progression to adenocarcinoma in Barrett’s esophagus. *Lab. Investigation* **86** 1261.
- SEGAL, D. M. and STEPHANY, D. A. (1984). The measurement of specific cell: Cell interactions by dual-parameter flow cytometry. *Cytometry A* **5** 169–181.
- SERTEL, O., KONG, J., CATALYUREK, U. V., LOZANSKI, G., SALTZ, J. H. and GURCAN, M. N. (2009a). Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. *J. Signal Processing Systems* **55** 169.
- SERTEL, O., KONG, J., SHIMADA, H., CATALYUREK, U. V., SALTZ, J. H. and GURCAN, M. N. (2009b). Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognit.* **42** 1093–1103.
- STOYAN, D. and PENTTINEN, A. (2000). Recent applications of point process methods in forestry statistics. *Statist. Sci.* **15** 61–78. [MR1842237](#)
- TABESH, A., TEVEROVSKIY, M., PANG, H.-Y., KUMAR, V. P., VERBEL, D., KOTSIANTI, A. and SAIDI, O. (2007). Multifeature prostate cancer diagnosis and Gleason grading of histological images. *IEEE Trans. Med. Imag.* **26** 1366–1378.
- TSAO, M.-S., MARGUET, S., LE TEUFF, G., LANTUEJOUL, S., SHEPHERD, F. A., SEYMOUR, L., KRATZKE, R., GRAZIANO, S. L., POPPER, H. H., ROSELL, R. et al. (2015). Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. *J. Clin. Oncol.* **33** 3439–3446.
- VAN LIESHOUT, M. N. M. and BADDELEY, A. J. (1996). A nonparametric measure of spatial interaction in point patterns. *Stat. Neerl.* **50** 344–361. [MR1422574](#)
- VAN LIESHOUT, M. N. M. and BADDELEY, A. J. (1999). Indices of dependence between types in multivariate point patterns. *Scand. J. Stat.* **26** 511–532. [MR1734259](#)
- VANEY, D. I., PEICHL, L. and BOYCOTT, B. (1981). Matching populations of amacrine cells in the inner nuclear and ganglion cell layers of the rabbit retina. *J. Comparative Neurology* **199** 373–391.
- VINCENT, L. and JEULIN, D. (1989). Minimal paths and crack propagation simulations. *Acta Stereologica* **8** 487–494.

- WALLER, L. A. (2005). Bayesian thinking in spatial statistics. In *Bayesian Thinking: Modeling and Computation. Handbook of Statist.* **25** 589–622. Elsevier/North-Holland, Amsterdam. [MR2490540](#)
- WANG, S., WANG, T., YANG, L., YI, F., LUO, X., YANG, Y., GAZDAR, A., FUJIMOTO, J., WISTUBA, I. I., YAO, B. et al. (2018). ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by convolutional neural network. Available at [arXiv:1809.10240](#).
- WÄSSLE, H. and ILLING, R.-B. (1981). Morphology and mosaic of on-and off-beta cells in the cat retina and some functional considerations. *Proc. Roy. Soc. London Ser. B, Biolog. Sci.* **212** 177–195.
- WÄSSLE, H., PEICHL, L. and BOYCOTT, B. B. (1981). Dendritic territories of cat retinal ganglion cells. *Nature* **292** 344–345.
- WÄSSLE, H. and RIEMANN, H. (1978). The mosaic of nerve cells in the mammalian retina. *Proc. Roy. Soc. London Ser. B, Biolog. Sci.* **200** 441–461.
- WIEGAND, T. and MOLONEY, K. A. (2004). Rings, circles, and null-models for point pattern analysis in ecology. *Oikos* **104** 209–229.
- WISEMAN, B. S. and WERB, Z. (2002). Stromal effects on mammary gland development and breast cancer. *Science* **296** 1046–1049.
- XIAO, G., REILLY, C. and KHODURSKY, A. B. (2009). Improved detection of differentially expressed genes through incorporation of gene locations. *Biometrics* **65** 805–814. [MR2649853](#)
- XIAO, G., WANG, X. and KHODURSKY, A. B. (2011). Modeling three-dimensional chromosome structures using gene expression data. *J. Amer. Statist. Assoc.* **106** 61–72. [MR2816702](#)
- YU, K.-H., ZHANG, C., BERRY, G. J., ALTMAN, R. B., RÉ, C., RUBIN, D. L. and SNYDER, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* **7**.
- YUAN, Y., FAILMEZGER, H., RUEDA, O. M., ALI, H. R., GRÄF, S., CHIN, S.-F., SCHWARZ, R. F., CURTIS, C., DUNNING, M. J., BARDWELL, H. et al. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Translational Med.* **4** 157ra143.

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

- BENAGLIA, T., CHAUVEAU, D., HUNTER, D. R. and YOUNG, D. (2009). mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32** 1–29.
- BRENNECKE, P., ANDERS, S., KIM, J. K., KOŁODZIEJCZYK, A. A., ZHANG, X., PROSERPIO, V., BAYING, B., BENES, V., TEICHMANN, S. A., MARIONI, J. C. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10** 1093–1095.
- CALIŃSKI, T. and HARABASZ, J. (1974). A dendrite method for cluster analysis. *Comm. Statist.* **3** 1–27. [MR0375641](#)
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. [MR0350949](#)
- FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCEL RATH, M. J., PRLIC, M. et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16** 278.
- FORGY, E. W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **21** 768–769.
- FRITSCH, A. and ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4** 367–391. [MR2507368](#)
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- LAKE, B. B., AI, R., KAESER, G. E., SALATHIA, N. S., YUNG, Y. C., LIU, R., WILDBERG, A., GAO, D., FUNG, H.-L., CHEN, S. et al. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352** 1586–1590.

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

- LIN, P., TROUP, M. and HO, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18** 59.
- LIU, Y., WARREN, J. L. and ZHAO, H. (2019). Supplement to “A hierarchical Bayesian model for single-cell clustering using RNA-sequencing data.” DOI:10.1214/19-AOAS1250SUPPA, DOI:10.1214/19-AOAS1250SUPPB.
- MAATEN, L. V. D. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
- MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman–Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. [MR3277163](#)
- MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. and WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5** 621–628.
- OKATY, B. W., FRERET, M. E., ROOD, B. D., BRUST, R. D., HENNESSY, M. L., KIM, J. C., COOK, M. N., DYMECKI, S. M. et al. (2015). Multi-scale molecular deconstruction of the serotonin neuron system. *Neuron* **88** 774–791.
- PATEL, A. P., TIROSH, I., TROMBETTA, J. J., SHALEK, A. K., GILLESPIE, S. M., WAKIMOTO, H., CAHILL, D. P., NAHED, B. V., CURRY, W. T., MARTUZA, R. L. et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344** 1396–1401.
- PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 241.
- POLLEN, A. A., NOWAKOWSKI, T. J., SHUGA, J., WANG, X., LEYRAT, A. A., LUI, J. H., LI, N., SZPANKOWSKI, L., FOWLER, B., CHEN, P. et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32** 1053–1058.
- RISSE, D. and COLE, M. (2016). scRNAseq: A collection of public single-cell RNA-seq datasets. R Package Version 1.4.0.
- STEGLE, O., TEICHMANN, S. A. and MARIONI, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16** 133–145.
- TASIC, B., MENON, V., NGUYEN, T. N., KIM, T. K., JARSKY, T., YAO, Z., LEVI, B., GRAY, L. T., SORESENSEN, S. A., DOLBEARE, T. et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19** 335–346.
- TSAFRIR, D., TSAFRIR, I., EIN-DOR, L., ZUK, O., NOTTERMAN, D. A. and DOMANY, E. (2005). Sorting points into neighborhoods (SPIN): Data analysis and visualization by ordering distance matrices. *Bioinformatics* **21** 2301–2308.
- USOSKIN, D., FURLAN, A., ISLAM, S., ABDO, H., LÖNNERBERG, P., LOU, D., HJERLING-LEFFLER, J., HAEGGSTRÖM, J., KHARCHENKO, O., KHARCHENKO, P. V. et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18** 145–153.
- VALLEJOS, C. A., RISSE, D., SCIALDONE, A., DUDOIT, S. and MARIONI, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14** 565–571.
- VINH, N. X., EPPS, J. and BAILEY, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11** 2837–2854. [MR2738784](#)
- WANG, B., ZHU, J., PIERSON, E., RAMAZZOTTI, D. and BATZOGLOU, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14** 414–416.
- XIE, F. and XU, Y. (2017). Bayesian repulsive Gaussian mixture model. ArXiv Preprint. Available at [arXiv:1703.09061](#).
- XU, C. and SU, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31** 1974–1980.

ZEISEL, A., MUÑOZ-MANCHADO, A. B., CODELUPPI, S., LÖNNERBERG, P., LA MANNO, G., JURÉUS, A., MARQUES, S., MUNGUBA, H., HE, L., BETSHOLTZ, C. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347** 1138–1142.

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 Fellegi 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

- ABRIL, D., NAVARRO-ARRIBAS, G. and TORRA, V. (2012). Improving record linkage with supervised learning for disclosure risk assessment. *Inf. Fusion* **13** 274–284.
- ALBERT, P. S. (2009). Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Stat. Med.* **28** 780–797. [MR2657042](#)
- ALBERT, P. S. and DODD, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60** 427–435. [MR2066277](#)
- ALBERT, P. S. and DODD, L. E. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *J. Amer. Statist. Assoc.* **103** 61–73. [MR2420217](#)

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

- ALBERT, P. S., MCSHANE, L. M. and SHIH, J. H. (2001). Latent class modelling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57** 610–619. [MR1855699](#)
- ARMSTRONG, J. B. and MAYDA, J. E. (1992). Estimation of record linkage models using dependent data. In *Proceedings of the Section on Survey Research Methods* 853–858. Amer. Statist. Assoc., Providence, RI.
- BILENKO, M. and MOONEY, R. J. (2003a). Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 39–48. ACM, New York.
- BILENKO, M. and MOONEY, R. J. (2003b). On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation* 7–12. Washington, DC.
- CHRISTEN, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 151–159. ACM, New York.
- CHRISTEN, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, Berlin.
- CLOGG, C. C. (1995). Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (G. Arminger, C. C. Clogg and S. ME, eds.) 311–359. Plenum, New York.
- COLLINS, J. and HUYNH, M. (2014). Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Stat. Med.* **33** 4141–4169. [MR3267401](#)
- COOK, L. J., OLSON, L. M. and DEAN, J. M. (2001). Probabilistic record linkage: Relationships between file sizes, identifiers and match weights. *Methods Inf. Med.* **40** 196–203.
- DAGGY, J., XU, H., HUI, S. L., GAMACHE, R. E. and GRANNIS, S. J. (2013). A practical approach for incorporating dependence among fields in probabilistic record linkage. *BMC Med. Inform. Decis. Mak.* **13** 97.
- DAGGY, J., XU, H., HUI, S. and GRANNIS, S. (2014). Evaluating latent class models with conditional dependence in record linkage. *Stat. Med.* **33** 4250–4265. [MR3267408](#)
- DENDUKURI, N. and JOSEPH, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57** 158–167. [MR1833302](#)
- ESPELAND, M. A. and HANDELMAN, S. L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45** 587–599.
- FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.
- FORTINI, M., LISEO, B., NUCCITELLI, A. and SCANU, M. (2001). On Bayesian record linkage. *Res. Off. Stat.* **4** 185–198.
- FORTINI, M., NUCCITELLI, A., LISEO, B. and SCANU, M. (2002). Modeling issues in record linkage: A Bayesian perspective. In *Proceedings of the Section on Survey Research Methods*. Amer. Statist. Assoc., Alexandria, VA.
- GARRETT, E. S. and ZEGER, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067. [MR1815583](#)
- GOMATAM, S., CARTER, R., ARIET, M. and MITCHELL, G. (2002). An empirical comparison of record linkage procedures. *Stat. Med.* **21** 1485–1496.
- GRANNIS, S. J., OVERHAGE, J. M. and McDONALD, C. J. (2004). Real world performance of approximate string comparators for use in patient matching. *Stud. Health Technol. Inform.* **107** 43–47.
- GRANNIS, S. J., OVERHAGE, J. M., HUI, S. L. and McDONALD, C. J. (2003). Analysis of a probabilistic record linkage technique without human review. *AMIA Annu. Symp. Proc.* **2003** 259–263.
- HADGU, A. and QU, Y. (1998). A biomedical application of latent class models with random effects. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 603–616.

- HAGENAARS, J. A. (1988). Latent structure models with direct effects between indicators, local dependence models. *Sociol. Methods Res.* **16** 379–405.
- HAN, H., GILES, L., ZHA, H., LI, C. and TSIOUTSIOLIKLIS, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* 296–305. ACM, New York.
- HAND, D. J. (2006). Classifier technology and the illusion of progress. *Statist. Sci.* **21** 1–34. [MR2275965](#)
- HAND, D. and YU, K. (2001). Idiot’s Bayes—Not so stupid after all? *Int. Stat. Rev.* **69** 385–398.
- HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, Berlin.
- KELLY, R. P. (1986). Robustness of the Census Bureau’s record linkage system. *Proc. Sect. Surv. Res. Methods* **1986** 620–624.
- LARSEN, M. D. (1997). Modeling issues and the use of experience in record linkage. In *Record Linkage Techniques—1997: Proceedings of an International Workshop and Exposition* (W. Alvey and B. Jamerson, eds.) 95–105. National Academies Press, Washington, DC.
- LARSEN, M. D. (2004). Record linkage using finite mixture models. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X.-L. Meng, eds.). *Wiley Ser. Probab. Stat.* 309–318. Wiley, Chichester. [MR2138266](#)
- LARSEN, M. D. (2012). An experiment with hierarchical Bayesian record linkage. Preprint. Available at [arxiv:1212.5203](#).
- LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. [MR1973781](#)
- LI, X. and SHEN, C. (2013). Linkage of patient records from disparate sources. *Stat. Methods Med. Res.* **22** 31–38. [MR3190644](#)
- LI, X., XU, H., SHEN, C. and GRANNIS, S. (2018). Automated linkage of patient records from disparate sources. *Stat. Methods Med. Res.* **27** 172–184. [MR3745662](#)
- MARTINS, B. (2011). A supervised machine learning approach for duplicate detection over gazetteer records. In *GeoSpatial Semantics: 4th International Conference, GeoS 2011, Brest, France, May 12–13, 2011. Proceedings* (C. Claramunt, S. Levashkin and M. Bertolotto, eds.) 34–51. Springer, Berlin.
- NEWCOMBE, H. B. and KENNEDY, J. M. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Commun. ACM* **5** 563–566.
- ONG, T. C., MANNINO, M. V., SCHILLING, L. M. and KAHN, M. G. (2014). Improving record linkage performance in the presence of missing linkage data. *J. Biomed. Inform.* **52** 43–54.
- QU, Y., TAN, M. and KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810. [MR1411731](#)
- QUANTIN, C., BINQUET, C., BOURGUARD, K., ALLAERT, F. A., FERDYNUS, C., PATTISINA, R., HARMENIL, G., PEGUIGNOT, S. and GOUYON, J. B. (2004). Assessment of the discriminating power of identifiers for record linkage. *Rev. épidémiol. Santé Publique* **52** 431–440.
- SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* **8** 2404–2434. [MR3292503](#)
- SADINLE, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *J. Amer. Statist. Assoc.* **112** 600–612. [MR3671755](#)
- SADINLE, M. and FIENBERG, S. E. (2013). A generalized Fellegi–Sunter framework for multiple record linkage with application to homicide record systems. *J. Amer. Statist. Assoc.* **108** 385–397. [MR3174628](#)
- SARIYAR, M., BORG, A. and POMMERENING, K. (2011). Controlling false match rates in record linkage using extreme value theory. *J. Biomed. Inform.* **44** 648–654.
- SCHMIDTMANN, I., HAMMER, G., SARIYAR, M. and GERHOLD-AY, A. (2009). Evaluation des Krebsregisters NRW Schwerpunkt record linkage. Technical report, In-

- stitute of Medical Biostatistics, Epidemiology and Informatics at Johannes Gutenberg Univ. Available at http://www.krebsregister.nrw.de/fileadmin/user_upload/dokumente/Evaluation/EKR_NRW_Evaluation_Abschlussbericht_2009-06-11.pdf.
- SEPÚLVEDA, R., VICENTE-VILLARDÓN, J. L. and GALINDO, M. P. (2008). The biplot as a diagnostic tool of local dependence in latent class models: A medical application. *Stat. Med.* **27** 1855–1869. [MR2420349](#)
- SUBTIL, A., DE OLIVEIRA, M. R. and GONÇALVES, L. (2012). Conditional dependence diagnostic in the latent class model: A simulation study. *Statist. Probab. Lett.* **82** 1407–1412. [MR2929794](#)
- TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. [MR2849786](#)
- THIBAudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Surv. Methodol.* **19** 31–38.
- TORRA, V., NAVARRO-ARRIBAS, G. and ABRIL, A. (2010). Supervised learning for record linkage through weighted means and OWA operators. *Control Cybernet.* **39** 1011–1026.
- TORRANCE-RYNARD, V. L. and WALTER, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Stat. Med.* **97** 2157–2175.
- TREERATPITUK, P. and GILES, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* 39–48. ACM, New York.
- TROMP, M., MERAY, N., RAVELLI, A. C. J., REITSMA, J. B. and BONSEL, G. J. (2008). Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies. *J. Am. Med. Inform. Assoc.* **15** 654–660.
- TROMP, M., RAVELLI, A. C. J., BONSEL, G. J., HASMAN, A. and REITSMA, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *J. Clin. Epidemiol.* **64** 565–572.
- UEBERSAX, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Appl. Psychol. Meas.* **23** 283–297.
- VACEK, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41** 959–968.
- VENTURA, S. L. and NUGENT, R. (2014). Hierarchical linkage clustering with distributions of distances for large-scale record linkage. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17–19, 2014. Proceedings* (J. Domingo-Ferrer, ed.). Springer, Berlin.
- VENTURA, S. L., NUGENT, R. and FUCHS, E. R. H. (2015). Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Res. Policy* **44** 1672–1701.
- WALTER, S. D., MACASKILL, P., LORD, S. J. and IRWIG, L. (2012). Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat. Med.* **31** 1129–1138. [MR2925684](#)
- WINKLER, W. (1988). Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 667–671. Amer. Statist. Assoc., Providence, RI.
- WINKLER, W. (1989). Methods for adjusting for lack of independence in an application of the Fellegi–Sunter model of record linkage. *Surv. Methodol.* **15** 101–117.
- WINKLER, W. (1993). Improved decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 274–279. Amer. Statist. Assoc., Providence.
- XU, H., BLACK, M. A. and CRAIG, B. A. (2013). Evaluating accuracy of diagnostic tests with intermediate results in the absence of a gold standard. *Stat. Med.* **32** 2571–2584. [MR3067408](#)
- XU, H. and CRAIG, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics* **65** 1145–1155. [MR2756502](#)

- XU, H., LI, X., SHEN, C., HUI, S. L. and GRANNIS, S. (2019). Supplement to “Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter?” DOI:[10.1214/19-AOAS1256SUPP](https://doi.org/10.1214/19-AOAS1256SUPP).
- YANG, I. and BECKER, B. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics* **53** 948–958.
- ZHU, R., ZHANG, J., ZHANG, D. and YAN, G. (2010). Stepwise variable selection for loglinear mixtures in record linkage. *Eur. J. Pure Appl. Math.* **3** 141–162. [MR2630106](#)

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

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. [MR0637937](#)
- BAKKOUR, A., MORRIS, J. C., WOLK, D. A. and DICKERSON, B. C. (2013). The effects of aging and Alzheimer's disease on cerebral cortical anatomy: Specificity and differential relationships with cognition. *NeuroImage* **76** 332–344.
- BONTHIUS, D. J., SOLODKIN, A. and VAN HOESSEN, G. W. (2005). Pathology of the insular cortex in Alzheimer disease depends on cortical architecture. *J. Neuropathol. Exp. Neurol.* **64** 910–922.

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

- CANU, E., McLAREN, D. G., FITZGERALD, M. E., BENDLIN, B. B., ZOCCATELLI, G., ALESSANDRINI, F., PIZZINI, F. B., RICCIARDI, G. K., BELTRAMELLO, A. et al. (2011). Mapping the structural brain changes in Alzheimer's disease: The independent contribution of two imaging modalities. *J. Alzheimer's Dis.* **26** 263–274.
- DAIANU, M., JAHANSHAD, N., NIR, T. M., JACK JR, C. R., WEINER, M. W., BERNSTEIN, M. A., THOMPSON, P. M. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2015). Rich club analysis in the Alzheimer's disease connectome reveals a relatively undisturbed structural core network. *Hum. Brain Mapp.* **36** 3087–3103.
- GAO, J., CHEUNG, R. T., CHAN, Y.-S., CHU, L.-W., MAK, H. K. and LEE, T. M. (2014). The relevance of short-range fibers to cognitive efficiency and brain activation in aging and dementia. *PLoS ONE* **9** e90307.
- GAO, Z., FU, H.-J., ZHAO, L.-B., SUN, Z.-Y., YANG, Y.-F. and ZHU, H.-Y. (2018). Aberrant DNA methylation associated with Alzheimer's disease in the superior temporal gyrus. *Exp. Ther. Med.* **15** 103–108.
- GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics **44**. Cambridge Univ. Press, Cambridge. [MR3587782](#)
- HOOVER, D. (1979). Relations on Probability Spaces and Arrays of Random Variables. Preprint, Institute for Advanced Study, Princeton, NJ.
- JACK, C. R., PETERSEN, R. C., XU, Y., O'BRIEN, P. C., SMITH, G. E., IVNIK, R. J., TANGALOS, E. G. and KOKMEN, E. (1998). Rate of medial temporal lobe atrophy in typical aging and Alzheimer's disease. *Neurology* **51** 993–999.
- KARAS, G., SCHELTENS, P., ROMBOUTS, S., VAN SCHIJNDEL, R., KLEIN, M., JONES, B., VAN DER FLIER, W., VRENKEN, H. and BARKHOF, F. (2007). Precuneus atrophy in early-onset Alzheimer's disease: A morphometric structural MRI study. *Neuroradiology* **49** 967–976.
- KARAVASILIS, E., PARTHIMOS, T. P., PAPATRIANTAFYLLOU, J. D., PAPAGEORGIOU, S. G., KAPASAS, G., PAPANICOLAOU, A. C. and SEIMENIS, I. (2017). A specific pattern of gray matter atrophy in Alzheimer's disease with depression. *J. Neurolinguist.* **264** 2101–2109.
- KLAASSENS, B. L., VAN GERVEN, J. M. A., VAN DER GROND, J., DE VOS, F., MÖLLER, C. and ROMBOUTS, S. A. R. B. (2017). Diminished posterior precuneus connectivity with the default mode network differentiates normal aging from Alzheimer's disease. *Front. Aging Neurosci.* **9** 97.
- LEECH, R. and SHARP, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain* **137** 12–32.
- LEHÉRICY, S., HIRSCH, E. C., HERSH, L. B. and AGID, Y. (1991). Cholinergic neuronal loss in the globus pallidus of Alzheimer disease patients. *Neurosci. Lett.* **123** 152–155.
- LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957. [MR2274085](#)
- MAECHLER, M. (2013). Package 'diptest'. In *R Package Version* 0.75–5.
- MCLIMANS, K. E. and WILLETTE, A. A. (2016). Novel CSF biomarker of metabolic dysfunction predicts AD-like associations across the Alzheimer's spectrum. *FASEB J.* **30** lb304–lb304.
- MIGLIACCIO, R., AGOSTA, F., POSSIN, K. L., CANU, E., FILIPPI, M., RABINOVICI, G. D., ROSEN, H. J., MILLER, B. L. and GORNO-TEMPINI, M. L. (2015). Mapping the progression of atrophy in early-and late-onset Alzheimer's disease. *J. Alzheimer's Dis.* **46** 351–364.
- PHILLIPS, D. J., MCGLAUGHLIN, A., RUTH, D., JAGER, L. R., SOLDAN, A., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2015). Graph theoretic analysis of structural connectivity across the spectrum of Alzheimer's disease: The importance of graph creation methods. *NeuroImage: Clinical* **7** 377–390.
- PRESCOTT, J. W., GUIDON, A., DORAISWAMY, P. M., ROY CHOUDHURY, K., LIU, C., PETRELLA, J. R. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2014). The

- Alzheimer structural connectome: Changes in cortical network topology with increased amyloid plaque burden. *Radiology* **273** 175–184.
- ROY, A., GHOSAL, S., PRESCOTT, J., CHOUDHURY, K. R. and FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2019). Supplement to "Bayesian Modeling of the Structural Connectome for Studying Alzheimer's Disease." DOI:[10.1214/19-AOAS1257SUPP](https://doi.org/10.1214/19-AOAS1257SUPP).
- SANZ-ARIGITA, E. J., SCHOONHEIM, M. M., DAMOISEAUX, J. S., ROMBOUTS, S. A. R. B., MARIS, E., BARKHOF, F., SCHELTENS, P. and STAM, C. J. (2010). Loss of 'small-world' networks in Alzheimer's disease: Graph analysis of fMRI resting-state functional connectivity. *PLoS ONE* **5** e13788.
- SCHEFF, S. W., PRICE, D. A., SCHMITT, F. A., SCHEFF, M. A. and MUFSON, E. J. (2011). Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's disease. *J. Alzheimer's Dis.* **24** 547–557.
- SHEN, W. and GHOSAL, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. [MR3426318](#)
- TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. [MR3640191](#)
- TUCHOLKA, A., GRAU-RIVERA, O., FALCON, C., RAMI, L., SANCHEZ-VALLE, R., LLADÓ, A., GISPert, J. D., MOLINUEVO, J. L., INITIATIVE, A. D. N. et al. (2018). Structural connectivity alterations along the Alzheimer's disease continuum: Reproducibility across two independent samples and correlation with cerebrospinal fluid amyloid- β and tau. *J. Alzheimer's Dis.* **61** 1575–1587.
- VAN BUUREN, S. and GROOTHUIS-OUUDSHOORN, K. (2010). Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 1–68.
- WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. [MR0909372](#)

WAVELET SPECTRAL TESTING: APPLICATION TO NONSTATIONARY CIRCADIAN RHYTHMS¹

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

- ATKINSON, A. D., HILL, R. R., PIGNATIELLO, J. J. JR., VINING, G. G., WHITE, E. D. and CHICKEN, E. (2017). Wavelet ANOVA approach to model validation. *Simul. Model. Pract. Theory* **78** 18–27.
- BELL-PEDERSEN, D., CASSONE, V. M., EARNEST, D. J., GOLDEN, S. S., HARDIN, P. E., THOMAS, T. L. and ZORAN, M. J. (2005). Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nat. Rev. Genet.* **6** 544–556.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- BUJDOSO, N. and DAVIS, S. J. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of *Arabidopsis thaliana*. *Front. Plant Sci.* **4** 3.
- CHO, H. (2016). A test for second-order stationarity of time series based on unsystematic subsamples. *Stat* **5** 262–277. [Paging previously given as 1–16]. [MR3584714](#)
- COSTA, M. J., FINKENSTÄDT, B., GOULD, P. D., FOREMAN, J., HALLIDAY, K. J., HALL, A. J. W. and RAND, D. A. (2011). Estimating periodicity of oscillatory time series through resampling techniques. Centre for Research in Statistical Methodology, Univ. Warwick.

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

- DAS, S. and NASON, G. P. (2016). Measuring the degree of non-stationarity of a time series. *Stat* **5** 295–305. [MR3589268](#)
- DUSIK, V., SENTHILAN, P. R., MENTZEL, B., HARTLIEB, H., WÜLBECK, C., YOSHII, T., RAABE, T. and HELFRICH-FÖRSTER, C. (2014). The MAP kinase p38 is part of *Drosophila melanogaster*'s circadian clock. *PLoS Genet.* **10** e1004565.
- EDWARDS, K. D., AKMAN, O. E., KNOX, K., LUMSDEN, P. J., THOMSON, A. W., BROWN, P. E., POKHILKO, A., KOZMA-BOGNAR, L., NAGY, F. et al. (2010). Quantitative analysis of regulatory flexibility under changing environmental conditions. *Mol. Syst. Biol.* **6** 424.
- FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. [MR1649196](#)
- FOLEY, J. A., DEFRIES, R., ASNER, G. P., BARFORD, C., BONAN, G., CARPENTER, S. R., CHAPIN, F. S., COE, M. T., DAILY, G. C. et al. (2005). Global consequences of land use. *Science* **309** 570–574.
- FRYZLEWICZ, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Stat.* **32** 503–528. [MR2112368](#)
- FRYZLEWICZ, P. and NASON, G. P. (2006). Haar–Fisz estimation of evolutionary wavelet spectra. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 611–634. [MR2301011](#)
- FRYZLEWICZ, P. and OMBAO, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *J. Amer. Statist. Assoc.* **104** 299–312. [MR2504379](#)
- FRYZLEWICZ, P., VAN BELLEGEM, S. and VON SACHS, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Statist. Math.* **55** 737–764. [MR2028615](#)
- GUO, W., DAI, M., OMBAO, H. C. and VON SACHS, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis. *J. Amer. Statist. Assoc.* **98** 643–652. [MR2011677](#)
- HARANG, R., BONNET, G. and PETZOLD, L. R. (2006). WAVOS: A MATLAB toolkit for wavelet analysis and visualization of oscillatory systems. *BMC Research Notes* **5** 163.
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J. and DAVIS, S. J. (2018). Clustering nonstationary circadian rhythms using locally stationary wavelet representations. *Multiscale Model. Simul.* **16** 184–214. [MR3749378](#)
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J., CHAWLA, S., MUNNS, J. and DAVIS, S. J. (2019). Supplement to “Wavelet spectral testing: application to nonstationary circadian rhythms.” DOI:10.1214/19-AOAS1246SUPP.
- HOLAN, S. H., WIKLE, C. K., SULLIVAN-BECKERS, L. E. and COCROFT, R. B. (2010). Modeling complex phenotypes: Generalized linear models using spectrogram predictors of animal communication signals. *Biometrics* **66** 914–924. [MR2758228](#)
- KILLICK, R., ECKLEY, I. A. and JONATHAN, P. (2013). A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electron. J. Stat.* **7** 1167–1183. [MR3056071](#)
- KON, N., SUGIYAMA, Y., YOSHITANE, H., KAMESHITA, I. and FUKADA, Y. (2015). Cell-based inhibitor screening identifies multiple protein kinases important for circadian clock oscillations. *Commun. Integr. Biol.* **8** e982405.
- KRZEMIENIEWSKA, K., ECKLEY, I. A. and FEARNHEAD, P. (2014). Classification of non-stationary time series. *Stat* **3** 144–157.
- LEISE, T. L., INDIC, P., PAUL, M. J. and SCHWARTZ, W. J. (2013). Wavelet meets actogram. *J. Biol. Rhythms* **28** 62–68.
- MARTINEZ, J. G., BOHN, K. M., CARROLL, R. J. and MORRIS, J. S. (2013). A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed models for nonstationary acoustic time series. *J. Amer. Statist. Assoc.* **108** 514–526. [MR3174638](#)
- MCCLUNG, C. R. (2006). Plant circadian rhythms. *Plant Cell* **18** 792–803.
- MCKAY, J. L., WELCH, T. D. J., VIDAČKOVIC, B. and TING, L. H. (2012). Statistically significant contrasts between EMG waveforms revealed using wavelet-based functional ANOVA. *J. Neurophysiol.* **109** 591–602.

- MILLAR, A. J., CARRINGTON, J. T., TEE, W. V. and HODGE, S. K. (2015). Changing planetary rotation rescues the biological clock mutant *lhy cca1* of *Arabidopsis thaliana*. *BioRxiv* at Cold Spring Harbor Laboratory.
- MINORS, D. S. and WATERHOUSE, J. M. (2013). *Circadian Rhythms and the Human*. Butterworth-Heinemann, Oxford.
- MOORE, D. S. (2007). *The Basic Practice of Statistics* **2**. WH Freeman, New York.
- MOORE, A., ZIELINSKI, T. and MILLAR, A. J. (2014). Online period estimation and determination of rhythmicity in circadian data, using the BioDare data infrastructure. *Methods Mol. Biol.* **1158** 13–44.
- MORRIS, J. S. (2015). Functional regression. *Ann. Rev. Stat. Appl.* **2** 321–359.
- MORRIS, J. S., BALADANDAYUTHAPANI, V., HERRICK, R. C., SANNA, P. and GUTSTEIN, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.* **5** 894–923. [MR2840180](#)
- NASON, G. P. (2008). *Wavelet Methods in Statistics with R. Use R!* Springer, New York. [MR2445580](#)
- NASON, G. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 879–904. [MR3124795](#)
- NASON, G. P. and SAVCHEV, D. (2014). White noise testing using wavelets. *Stat* **3** 351–362.
- NASON, G. P. and STEVENS, K. (2015). Bayesian wavelet shrinkage of the Haar–Fisz transformed wavelet periodogram. *PLoS ONE* **10** e0137662.
- NASON, G. P., VON SACHS, R. and KROISANDT, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 271–292. [MR1749539](#)
- NICHOLSON, F. A., SMITH, S. R., ALLOWAY, B. J., CARLTON-SMITH, C. and CHAMBERS, B. J. (2003). An inventory of heavy metals inputs to agricultural soils in England and Wales. *Sci. Total Environ.* **311** 205–219.
- OGDEN, T. R. (1997). On preconditioning the data for the wavelet transform when the sample size is not a power of two. *Comm. Statist. Simulation Comput.* **26** 467–486.
- OH, H.-S., AMMANN, C. M., NAVEAU, P., NYCHKA, D. and OTTO-BLIESNER, B. L. (2003). Multi-resolution time series analysis applied to solar irradiance and climate reconstructions. *J. Atmos. Sol.-Terr. Phys.* **65** 191–201.
- PEREA-GARCÍA, A., ANDRÉS-BORDERÍA, A., DE ANDRÉS, S. M., SANZ, A., DAVIS, A. M., DAVIS, S. J., HUIJSER, P. and PEÑARRUBIA, L. (2015). Modulation of copper deficiency responses by diurnal and circadian rhythms in *arabidopsis thaliana*. *J. Exp. Bot.* **67** 391–403.
- PRICE, T. S., BAGGS, J. E., CURTIS, A. M., FITZGERALD, G. A. and HOGENESCH, J. B. (2008). WAVECLOCK: Wavelet analysis of circadian oscillation. *Bioinformatics* **24** 2794–2795.
- PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series. Vol. 1. Univariate Series, Probability and Mathematical Statistics*. Academic Press, New York. [MR0628735](#)
- PRIESTLEY, M. B. and SUBBA RAO, T. (1969). A test for non-stationarity of time-series. *J. Roy. Statist. Soc. Ser. B* **31** 140–149. [MR0269062](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](#)
- SENEŠIL, G. S., BALDASSARRE, G., SENESI, N. and RADINA, B. (1998). Trace element inputs into soils by anthropogenic activities and implications for human health. *Chemosphere* **39** 343–377.
- SHUMWAY, R. H. (1988). *Applied Statistical Time Series Analysis*. Prentice-Hall, Englewood Cliffs, NJ.
- TAVAKOLI, S. and PANARETOS, V. M. (2016). Detecting and localizing differences in functional time series dynamics: A case study in molecular biophysics. *J. Amer. Statist. Assoc.* **111** 1020–1035. [MR3561926](#)

- TORRENCE, C. and COMPO, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79** 61–78.
- VAN BELLEGEM, S. and VON SACHS, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Ann. Statist.* **36** 1879–1924. [MR2435459](#)
- VIDAKOVIC, B. (2001). Wavelet-based functional data analysis: Theory, applications and ramifications. In *Proceedings of the 3rd Pacific Symposium on Flow Visualization and Image Processing*, Maui, HI.
- VITATERNA, M. H., TAKAHASHI, J. S. and TUREK, F. W. (2001). Overview of circadian rhythms. *Alcohol Res. Health* **25** 85–93.
- VON SACHS, R. and NEUMANN, M. H. (2000). A wavelet-based test for stationarity. *J. Time Series Anal.* **21** 597–613. [MR1794489](#)
- ZIELINSKI, T., MOORE, A. M., TROUP, E., HALLIDAY, K. J. and MILLAR, A. J. (2014). Strengths and limitations of period estimation methods for circadian data. *PLoS ONE* **9** 96462.

OBLIQUE RANDOM SURVIVAL FORESTS¹

BY 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[§]

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

- ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (2012). *Statistical Models Based on Counting Processes*. Springer, Berlin
- BIEN, J. and TIBSHIRANI, R. (2019). *protoclust*: Hierarchical clustering with prototypes. R package version 1.6.3.
- BINDER, H. (2013). *CoxBoost*: Cox models by likelihood based boosting for a single survival endpoint or competing risks. R package version 1.4, available at <https://CRAN.R-project.org/package=CoxBoost>.
- BLANCHE, P., KATTAN, M. W. and GERDS, T. A. (2019). The c-index is not proper for the evaluation of *t*-year predicted risks. *Biostatistics* **20** 347–357. [MR3922138](#)
- BOU-HAMAD, I., LAROCQUE, D. and BEN-AMEUR, H. (2011). A review of survival trees. *Stat. Surv.* **5** 44–71. [MR3018509](#)
- BREIMAN, L. (1984). *Classification and Regression Trees*. Routledge, Abingdon.
- BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

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

- BREIMAN, L. and CUTLER, A. (2003). Setting up, using, and understanding random forests V4.0. Dept. Statistics, Univ. California, Berkeley.
- BRILLEMAN, S. (2018). *simSurv*: Simulate survival data. R package version 0.2.2, available at <https://CRAN.R-project.org/package=simSurv>.
- BURNHAM, K. P. and ANDERSON, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33** 261–304. [MR2086350](#)
- CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining* 785–794. ACM.
- CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I. et al. (2019). xgboost: Extreme gradient boosting. R package version 0.81.0.1, available at <https://CRAN.R-project.org/package=xgboost>.
- COX, D. R. (1992). Regression models and life-tables. In *Breakthroughs in Statistics* 527–541. Springer, Berlin.
- DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7** 1–30. [MR2274360](#)
- DESMEDT, C., DI LEO, A., DE AZAMBUJA, E., LARSIMONT, D., HAIBE-KAINS, B., SELLESLAGS, J., DELALOGUE, S., DUHEM, C., KAINS, J.-P. et al. (2011). Multifactorial approach to predicting resistance to anthracyclines. *J. Clin. Oncol.* **29** 1578–1586.
- DHEERU, D. and KARRA TANISKIDOU, E. (2017). UCI Machine learning repository. Univ. California, Irvine.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. [MR2060166](#)
- FLEMING, T. R. and HARRINGTON, D. P. (2011). *Counting Processes and Survival Analysis* **169**. Wiley, New York.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. [MR1873328](#)
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22. Available at <http://www.jstatsoft.org/v33/i01/>.
- FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–701.
- GERDS, T. A., KATTAN, M. W., SCHUMACHER, M. and YU, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat. Med.* **32** 2173–2184. [MR3067376](#)
- GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Mach. Learn.* **63** 3–42.
- GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18** 2529–2545.
- HARRELL, F. E., CALIFF, R. M., PRYOR, D. B., LEE, K. L. and ROSATI, R. A. (1982). Evaluating the yield of medical tests. *JAMA* **247** 2543–2546.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer Series in Statistics*. Springer, New York. [MR1851606](#)
- HATZIS, C., PUSZTAI, L., VALERO, V., BOOSER, D. J., ESSERMAN, L., LLUCH, A., VIDAURRE, T., HOLMES, F., SOUCHON, E. et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305** 1873–1881.
- HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344.
- HEAGERTY, P. J. and ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61** 92–105. [MR2135849](#)

- HOTHORN, T., HORNIK, K., STROBL, C. and ZEILEIS, A. (2019). party: A laboratory for recursive partytioning. R package version 1.3.3, available at <https://CRAN.R-project.org/package=party>.
- HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. [MR2291267](#)
- HOTHORN, T. and LAUSEN, B. (2003). Double-bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognit.* **36** 1303–1309.
- HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Stat. Med.* **23** 77–91.
- HOWARD, V. J., CUSHMAN, M., PULLEY, L., GOMEZ, C. R., GO, R. C., PRINEAS, R. J., GRAHAM, A., MOY, C. S. and HOWARD, G. (2005). The reasons for geographic and racial differences in stroke study: Objectives and design. *Neuroepidemiology* **25** 135–143.
- IMAN, R. L. and DAVENPORT, J. M. (1980). Approximations of the critical region of the Fbietkan statistic. *Comm. Statist. Theory Methods* **9** 571–595.
- ISHWARAN, H. and KOGALUR, U. B. (2019). Random forests for survival, regression, and classification (RF-SRC). R package version 2.8.0, available at <https://cran.r-project.org/package=randomForestSRC>.
- ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. [MR2516796](#)
- JAEGER, B. (2018). obliqueRSF: Oblique random forests for right-censored time-to-event data. R package version 0.1.0, available at <https://CRAN.R-project.org/package=obliqueRSF>.
- JAEGER, B. C., LONG, L. D., LONG, D. M., SIMS, M., SZYCHOWSKI, J. M., MIN, Y.-I., MCCLURE, L. A., HOWARD, G. and SIMON, N. (2019). Supplement to “Oblique random survival forests.” DOI:[10.1214/19-AOAS1261SUPP](#).
- KOWARIK, A. and TEMPL, M. (2016). Imputation with the R package VIM. *J. Stat. Softw.* **74** 1–16.
- LEVY, A. S., STEVENS, L. A., SCHMID, C. H., ZHANG, Y. L., CASTRO, A. F., FELDMAN, H. I., KUSEK, J. W., EGGERS, P., VAN LENTE, F. et al. (2009). A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* **150** 604–612.
- LUNDBERG, S. M., ERION, G. G. and LEE, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint [arXiv:1802.03888](#).
- MCCALL, M. N., BOLSTAD, B. M. and IRIZARRY, R. A. (2010). Frozen robust multiarray analysis (fRMA). *Biostatistics* **11** 242–253.
- MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41. [MR3491120](#)
- MENZE, B. H., KELM, B. M., SPLITTHOFF, D. N., KOETHE, U. and HAMPRECHT, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 453–469. Springer, Berlin.
- MOGENSEN, U. B., ISHWARAN, H. and GERDS, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50** 1.
- MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. [MR3937487](#)
- NASEJJE, J. B., MWAMBI, H., DHEDA, K. and LESOSKY, M. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med. Res. Methodol.* **17** 115.
- RAINFORTH, T. and WOOD, F. (2015). Canonical correlation forests. arXiv preprint [arXiv:1507.05444](#).
- SAFFORD, M. M., BROWN, T. M., MUNTNER, P. M., DURANT, R. W., GLASSER, S., HALANYCH, J. H., SHIKANY, J. M., PRINEAS, R. J., SAMDARSHI, T. et al. (2012). Association of race and sex with risk of incident acute coronary heart disease events. *JAMA* **308** 1768–1774.
- SCHUMACHER, M., BASTERT, G., BOJAR, H., HUEBNER, K., OLSCHIEWSKI, M., SAUERBREI, W., SCHMOOR, C., BEYERLE, C., NEUMANN, R. et al. (1994). Randomized 2 × 2 trial

- evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J. Clin. Oncol.* **12** 2086–2093.
- SEGAL, M. R. (1988). Regression trees for censored data. *Biometrics* **44** 35–47.
- SHABALIN, A. A., TJELMELAND, H., FAN, C., PEROU, C. M. and NOBEL, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* **24** 1154–1160.
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13.
- STRASSER, H. and WEBER, C. (1999). The asymptotic theory of permutation statistics. *Math. Methods Statist.* **8** 220–250. Johann Pfanzagl—on the occasion of his 70th birthday. [MR1722622](#)
- STROBL, C., MALLEY, J. and TUTZ, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **14** 323–348.
- STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. and HOTHORN, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **8** 25.
- TAYLOR JR., H. A., WILSON, J. G., JONES, D. W., SARPONG, D. F., SRINIVASAN, A., GARRISON, R. J., NELSON, C. and WYATT, S. B. (2005). Toward resolution of cardiovascular health disparities in African americans: Design and methods of the Jackson heart study. *Ethn. Dis.* **15** S6–4.
- TERNÈS, N., ROTOLO, F., HEINZE, G. and MICHIELS, S. (2017). Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biom. J.* **59** 685–701. [MR3672690](#)
- THERNEAU, T. M. (2015). A package for survival analysis in S. R package version 2.38, available at <https://CRAN.R-project.org/package=survival>.
- TUTZ, G. and BINDER, H. (2007). Boosting ridge regression. *Comput. Statist. Data Anal.* **51** 6044–6059. [MR2407697](#)
- VAN HOUWELINGEN, H. C., BRUINSMA, T., HART, A. A. M., VAN’T VEER, L. J. and WESSELS, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Stat. Med.* **25** 3201–3216. [MR2252292](#)
- VAN’T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y. D., HART, A. A., MAO, M., PETERSE, H. L., VAN DER KOOY, K., MARTON, M. J. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** 530.
- VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.
- WHELTON, P. K., CAREY, R. M., ARONOW, W. S., CASEY, D. E., COLLINS, K. J., HIMMELFARB, C. D., DEPALMA, S. M., GIDDING, S., JAMERSON, K. A. et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American college of cardiology/American heart association task force on clinical practice guidelines. *J. Am. Coll. Cardiol.* **71** e127–e248.
- ZHU, R. (2013). *Tree-Based Methods for Survival Analysis and High-Dimensional Data*. Thesis (Ph.D.)—Univ. North Carolina at Chapel Hill. ProQuest LLC, Ann Arbor, MI. [MR3211552](#)
- ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc.* **110** 1770–1784. [MR3449072](#)
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327](#)

APPROXIMATE INFERENCE FOR CONSTRUCTING ASTRONOMICAL CATALOGS FROM IMAGES¹

BY JEFFREY REGIER^{*}, ANDREW C. MILLER[†], DAVID SCHLEGEL[‡],
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

- BARBARY, K. (2016). SEP: Source extractor as a library. *The Journal of Open Source Software* **1** 59.
- BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR0804611](#)
- BERTIN, E. and ARNOUTS, S. (1996). SExtractor: Software for source extraction. *Astron. Astrophys. Suppl. Ser.* **117** 393–404.
- BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. [MR3605826](#)
- BICKEL, P. J. and DOKSUM, K. A. (2016). *Mathematical Statistics—Basic Ideas and Selected Topics. Vol. 2*, 2nd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3287337](#)
- BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, New York. [MR2247587](#)
- BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. [MR3671776](#)
- BOOTH, J. G. and SARKAR, S. (1998). Monte Carlo approximation of bootstrap variances. *Amer. Statist.* **52** 354–357.
- BOSCH, J., ARMSTRONG, R., BICKERTON, S. et al. (2018). The Hyper Suprime-Cam software pipeline. *Publ. Astron. Soc. Jpn.* **70**.
- BREWER, B. J., FOREMAN-MACKEY, D. and HOGG, D. W. (2013). Probabilistic catalogs for crowded stellar fields. *Astron. J.* **146**.

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

- BUBECK, S. (2015). Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.* **8**.
- DOI, M., TANAKA, M. et al. (2010). Photometric response functions of the Sloan Digital Sky Survey imager. *Astron. J.* **139**.
- EISENSTEIN, D. J., ANNIS, J., GUNN, J. E. et al. (2001). Spectroscopic target selection for the Sloan Digital Sky Survey: The Luminous red galaxy sample. *Astron. J.* **122** 2267.
- FAN, Y. and SISSON, S. A. (2011). Reversible Jump Markov Chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*.
- FEIGELSON, E. D. and BABU, G. J. (2012). *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge Univ. Press, Cambridge. [MR2963292](#)
- FELDMAN, M. (2018). Summit up and running at Oak Ridge. Available at <https://www.top500.org/news/summit-up-and-running-at-oak-ridge-claims-first-exascale-application/>. [Online; accessed August 20, 2018].
- FOLK, M., HEBER, G., KOZIOL, Q., POURMAL, E. and ROBINSON, D. (2011). An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases* 36–47.
- FRAYSSE, A. and RODET, T. (2014). A measure-theoretic variational Bayesian algorithm for large dimensional problems. *SIAM J. Imaging Sci.* **7** 2591–2622. [MR3285234](#)
- FRIEL, N. and WYSE, J. (2012). Estimating the evidence—a review. *Stat. Neerl.* **66** 288–308. [MR2955421](#)
- FUKUGITA, M., ICHIKAWA, T., GUNN, J. E., DOI, M., SHIMASAKU, K. and SCHNEIDER, D. P. (1996). The Sloan Digital Sky Survey photometric system. *Astron. J.* **111**.
- GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. [MR3235677](#)
- GIORDANO, R. J., BRODERICK, T. and JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems* 1441–1449.
- GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. [MR1380810](#)
- HAGER, G. and WELLEIN, G. (2010). *Introduction to High Performance Computing for Scientists and Engineers*. CRC Press, Boca Raton, FL.
- HERSHEY, J. R. and OLSEN, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing* **4** 317–320.
- JULIA DEVELOPERS (2018). Julia micro-benchmarks. Available at <https://julialang.org/benchmarks/>. [Online; accessed January 23, 2018].
- LANG, D., HOGG, D. W. and MYKYTYN, D. (2016). The Tractor: Probabilistic astronomical source detection and measurement. Astrophysics Source Code Library.
- LSST (2017). About LSST. Available at <http://www.lsst.org/about>. [Online; accessed September 12, 2017].
- LUPTON, R. H., IVEZIC, Z. et al. (2005). SDSS image processing II: The photo pipelines. Technical Report, Princeton Univ. Preprint. Available at <https://www.astro.princeton.edu/~rhl/photo-lite.pdf>.
- LUPTON, R., GUNN, J. E., IVEZIĆ, Z., KNAPP, G. R. and KENT, S. (2001). The SDSS imaging pipelines. In *Astronomical Data Analysis Software and Systems X* (F. R. Harnden, Jr., F. A. Primini and H. E. Payne, eds.). *Astronomical Society of the Pacific Conference Series* **238** 269–280.
- MACKAY, D. J. (1995). Developments in probabilistic modelling with neural networks—ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications*. Springer, Berlin.

- MELCHIOR, P., MOOLEKAMP, F., JERDEE, M., ARMSTRONG, R., SUN, A.-L., BOSCH, J. and LUPTON, R. (2018). SCARLET: Source separation in multi-band images by constrained matrix factorization. Available at [arXiv:1802.10157](https://arxiv.org/abs/1802.10157).
- MORGAN, T. (2018). The end of Xeon Phi. Available at <https://www.nextplatform.com/2018/07/27/end-of-the-line-for-xeon-phi-its-all-xeon-from-here/>. [Online; accessed August 20, 2018].
- MUJTABA, H. (2018). Intel Xeon Scalable Family Roadmap Revealed. Available at <https://wccftech.com/intel-xeon-scalable-family-roadmap-revealed-points-out-cascade-lake-sp-in-q4-2018-cooper-lake-sp-in-q4-2019-ice-lake-sp-in-1h-2020/>. [Online; accessed August 20, 2018].
- MURPHY, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.
- NEAL, R. M. (2001). Annealed importance sampling. *Stat. Comput.* **11** 125–139. [MR1837132](#)
- NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. [MR1994729](#)
- NERSC (2018). Cori configuration. Available at <http://www.nersc.gov/users/computational-systems/cori/configuration/>. [Online; accessed January 23, 2018].
- NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. [MR2244940](#)
- PORTILLO, S. K. N., LEE, B. C. G., DAYLAN, T. and FINKBEINER, D. P. (2017). Improved point-source detection in crowded fields using probabilistic cataloging. *Astron. J.* **154**.
- PRICE, D. C., BARSDELL, B. R. and GREENHILL, L. J. (2015). HDFITS: Porting the FITS data model to HDF5. *Astron. Comput.* **12**.
- REGIER, J., MCAULIFFE, J. and PRABHAT (2015). A deep generative model for astronomical images of galaxies. In *NIPS Workshop: Advances in Approximate Bayesian Inference*.
- REGIER, J., MILLER, A. C., SCHLEGEL, D., ADAMS, R. P., MCAULIFFE, J. D. and PRABHAT (2019). Supplement to “Approximate inference for constructing astronomical catalogs from images.” DOI:[10.1214/19-AOAS1258SUPP](https://doi.org/10.1214/19-AOAS1258SUPP).
- RICHARDS, G. T., FAN, X., NEWBERG, H. J., STRAUSS, M. A. et al. (2002). Spectroscopic target selection in the Sloan Digital Sky Survey: The quasar sample. *Astron. J.* **123** 2945.
- ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. [MR1707311](#)
- ROWE, B., JARVIS, M., MANDELBAUM, R. et al. (2015). GALSIM: The modular galaxy image simulation toolkit. *Astron. Comput.* **10**.
- SCHMIDT, M. N., WINTHER, O. and HANSEN, L. K. (2009). Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation* 540–547.
- SDSS (2017). Measures of flux and magnitude. Available at <http://www.sdss3.org/dr8/algorithms/magnitudes.php>. [Online; accessed November 12, 2017].
- SDSS (2018a). Glossary of SDSS-IV terminology. Available at <https://www.sdss.org/dr14/help/glossary/#N>. [Online; accessed August 21, 2018].
- SDSS (2018b). Camera. Available at <http://www.sdss.org/instruments/camera/>. [Online; accessed January 30, 2018].
- SDSS (2018c). Sky coverage. Available at <http://classic.sdss.org/dr7/coverage/>. [Online; accessed January 30, 2018].
- SDSS (2018d). Data model: Frame. https://data.sdss.org/datamodel/files/BOSS_PHOTOOBJ/frames/RERUN/RUN/CAMCOL/frame.html. [Online; accessed August 21, 2018].
- SKILLING, J. (2004). Nested sampling. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. *AIP Conf. Proc.* **735** 395–405. Amer. Inst. Phys., Melville, NY. [MR2266273](#)
- ŠMÍDL, V. and QUINN, A. (2008). Variational Bayesian filtering. *IEEE Trans. Signal Process.* **56** 5020–5030. [MR2517232](#)
- SMITH, R. (2018). AMD reaffirms 7nm Epyc Rome server processors sampling in 2H 2018. Available at <https://www.anandtech.com/show/13122/amd-rome-epyc-cpus-to-be-fabbed-by-tsmc>. [Online; accessed August 20, 2018].

- STRAUSS, M. A., WEINBERG, D. H., LUPTON, R. H. et al. (2002). Spectroscopic target selection in the Sloan Digital Sky Survey: The main galaxy sample. *Astron. J.* **124** 1810.
- TOP500.ORG (2017). Top500 List—November 2017. Available at <https://www.top500.org/list/2017/11/>. [Online; accessed November 16, 2017].
- TURON, C., MEYNADIER, F., ARENOU, F., HOGG, D. and LANG, D. (2010). *Telescopes Don't Make Catalogues! European Astronomical Society Publications Series* **45**.
- VAN LEEUWEN, D. (2018). GaussianMixtures.jl. Available at <https://github.com/davidavdav/GaussianMixtures.jl>. [Online; accessed August 21, 2018].
- WELLS, D. C. and GREISEN, E. W. (1979). FITS—a flexible image transport system. In *Image Processing in Astronomy* 445–471.
- WILSON, H. (2009). M33—“The Galaxy in Triangulum.” Available at <https://hwilson.zenfolio.com/galaxies/h274E3B43#h274e3b43>.
- YORK, D. G., ADELMAN, J., ANDERSON, J., ANDERSON, S. F., ANNIS, J., BAHCALL, N. A., BAKKEN, J. A. et al. (2000). The Sloan Digital Sky Survey: Technical summary. *Astron. J.* **120** 1579–1587.
- ZHENG, Y., FRAYSSE, A. and RODET, T. (2015). Efficient variational Bayesian approximation method based on subspace optimization. *IEEE Trans. Image Process.* **24** 681–693. [MR3301262](#)

BAYESIAN METHODS FOR MULTIPLE MEDIATORS: RELATING PRINCIPAL STRATIFICATION AND CAUSAL MEDIATION IN THE ANALYSIS OF POWER PLANT EMISSION CONTROLS¹

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

- ALBERT, J. M. and NELSON, S. (2011). Generalized causal mediation analysis. *Biometrics* **67** 1028–1038. [MR2829237](#)
- BACCINI, M., MATTEI, A. and MEALLI, F. (2015). Bayesian inference for causal mechanisms with application to a randomized study for postoperative pain control. DISIA Working Paper.
- BARON, R. M. and KENNY, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

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

- BARTOLUCCI, F. and GRILLI, L. (2011). Modeling partial compliance through copulas in a principal stratification framework. *J. Amer. Statist. Assoc.* **106** 469–479. [MR2866975](#)
- CHESTNUT, L. G. and MILLS, D. M. (2005). A fresh look at the benefits and costs of the US acid rain program. *J. Environ. Manag.* **77** 252–266.
- CONLON, A. S. C., TAYLOR, J. M. G. and ELLIOTT, M. R. (2017). Surrogacy assessment using principal stratification and a Gaussian copula model. *Stat. Methods Med. Res.* **26** 88–107. [MR3592714](#)
- DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N. and VANSTEELANDT, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71** 1–14. [MR3335344](#)
- DANIELS, M. J., ROY, J. A., KIM, C., HOGAN, J. W. and PERRI, M. G. (2012). Bayesian inference for the causal effect of mediation. *Biometrics* **68** 1028–1036. [MR3040009](#)
- DOMINICI, F., GREENSTONE, M. and SUNSTEIN, C. R. (2014). Particulate matter matters. *Science* **344** 257–259. DOI:[10.1126/science.1247348](#).
- FORASTIERE, L., MEALLI, F. and VANDERWEELE, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *J. Amer. Statist. Assoc.* **111** 510–525. [MR3538683](#)
- FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. [MR1891039](#)
- GILBERT, P. B. and HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. [MR2522262](#)
- GUSTAFSON, P. (2010). Bayesian inference for partially identified models. *Int. J. Biostat.* **6** Art. 17, 20. [MR2602560](#)
- HEI ACCOUNTABILITY WORKING GROUP (2003). *Assessing the Health Impact of Air Quality Regulations: Concepts and Methods for Accountability Research*. Communication 11. Health Effects Institute, Boston, MA.
- HODAN, W. M. and BARNARD, W. R. (2004). Evaluating the contribution of PM_{2.5} precursor gases and re-entrained road emissions to mobile source PM_{2.5} particulate matter emissions. In *13th International Emission Inventory Conference “Working for Clean Air in Clearwater.”*
- IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. [MR2741814](#)
- IMAI, K. and YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanism: Revisiting evidence from framing experiments. *Polit. Anal.* **21** 141–171.
- JOFFE, M. M. and GREENE, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538. [MR2751477](#)
- KIM, C., DANIELS, M. J., MARCUS, B. H. and ROY, J. A. (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73** 401–409. [MR3665957](#)
- KIM, C., DANIELS, M. J., HOGAN, J. W., CHOIRAT, C. and ZIGLER, C. M. (2019). Supplement to “Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls.” DOI:[10.1214/19-AOAS1260SUPP](#).
- MA, Y., ROY, J. and MARCUS, B. (2011). Causal models for randomized trials with two active treatments and continuous compliance. *Stat. Med.* **30** 2349–2362. [MR2829137](#)
- MACKINNON, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Earlbaum Associates, New York.
- MATTEI, A. and MEALLI, F. (2011). Augmented designs to assess principal strata direct effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 729–752. [MR2867456](#)
- MEALLI, F. and MATTEI, A. (2012). A refreshing account of principal stratification. *Int. J. Biostat.* **8** Art. 8, 19. [MR2923283](#)
- MEALLI, F. and PACINI, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131. [MR3174688](#)
- MEALLI, F. and RUBIN, D. B. (2003). Assumptions allowing the estimation of direct causal effects. *J. Econometrics* **112** 79–87. [MR1963228](#)

- MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. [MR1399156](#)
- NELSEN, R. B. (1999). *An Introduction to Copulas. Lecture Notes in Statistics* **139**. Springer, New York. [MR1653203](#)
- OFFICE OF MANAGEMENT AND BUDGET (2013). 2013 Draft report to Congress on the benefits and costs of federal regulation and unfunded mandates on state, local, and tribal entities. Technical report, OMB, Washington, DC.
- PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* 411–420. Morgan Kaufman, San Francisco, CA.
- POPE III, C. A., EZZATI, M. and DOCKERY, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. *N. Engl. J. Med.* **360** 376–386.
- ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **58** 688–701.
- RUBIN, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scand. J. Stat.* **31** 161–170. [MR2066246](#)
- SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **106** 1331–1344. [MR2896839](#)
- TAGURI, M., FEATHERSTONE, J. and CHENG, J. (2018). Causal mediation analysis with multiple causally non-ordered mediators. *Stat. Methods Med. Res.* **27** 3–19. [MR3745651](#)
- 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*.
- VANDERWEELE, T. J. (2008). Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.* **78** 2957–2962. [MR2516810](#)
- VANDERWEELE, T. J. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.
- VANDERWEELE, T. J. (2011). Principal stratification—uses and limitations. *Int. J. Biostat.* **7** Art. 28, 14. [MR2843532](#)
- VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiol Methods* **2** 95–115.
- WANG, W., NELSON, S. and ALBERT, J. M. (2013). Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. *Stat. Med.* **32** 4211–4228. [MR3118350](#)
- ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics (Oxford, England)* **13** 289–302. DOI:[10.1093/biostatistics/kxr052](#).

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

BY YOUYI ZHANG^{*}, JEFFREY S. MORRIS^{*,2}, SHIVALI NARANG AERRY[†],
 ARVIND U. K. RAO^{‡,3} 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

- AERTS, H. J., VELAZQUEZ, E. R., LEIJENAAAR, R. T., PARMAR, C., GROSSMANN, P., CARVALHO, S., BUSSINK, J., MONSHOUWER, R., HAIBE-KAINS, B., RIETVELD, D. et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**.
- ARMAGAN, A., DUNSON, D. B. and LEE, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23** 119–143. [MR3076161](#)

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

- AZADEH, S., HOBBS, B. P., MA, L., NIELSEN, D. A., MOELLER, F. G. and BALADANDAYUTHAPANI, V. (2016). Integrative Bayesian analysis of neuroimaging-genetic data with application to cocaine dependence. *NeuroImage* **125** 813–824.
- BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. [MR2065192](#)
- BATMANGHELICH, N. K., DALCA, A. V., SABUNCU, M. R. and GOLLAND, P. (2013). Joint modeling of imaging and genetics. In *Information Processing in Medical Imaging* 766–777. Springer, Berlin.
- BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. [MR3449048](#)
- BLEEKER, F. E., MOLENAAR, R. J. and LEENSTRA, S. (2012). Recent advances in the molecular understanding of glioblastoma. *J. Neurooncol.* **108** 11–27.
- CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. [MR2650751](#)
- CASTELLANO, G., BONILHA, L., LI, L. and CENDES, F. (2004). Texture analysis of medical images. *Clin. Radiol.* **59** 1061–1069.
- CHADDAD, A. and TANOUGAST, C. (2016). Extracted magnetic resonance texture features discriminate between phenotypes and are associated with overall survival in glioblastoma multiforme patients. *Med. Biol. Eng. Comput.* 1–12.
- COROLLER, T. P., GROSSMANN, P., HOU, Y., VELAZQUEZ, E. R., LEIJENAAR, R. T., HERMANN, G., LAMBIN, P., HAIBE-KAINS, B., MAK, R. H. et al. (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114** 345–350.
- DAEMEN, A., GEVAERT, O., OJEDA, F., DEBUCQUOY, A., SUYKENS, J. A., SEMPoux, C., MACHIELS, J.-P., HAUSERMANS, K. and MOOR, B. D. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Gen. Med.* **1** 39.
- FELIPE DE SOUSA, E. M., VERMEULEN, L., FESSLER, E. and MEDEMA, J. P. (2013). Cancer heterogeneity—a multifaceted view. *EMBO Rep.* **14** 686–695.
- GANESHAN, B., ABALKE, S., YOUNG, R., CHATWIN, C. R. and MILES, K. A. (2010). Texture analysis of non-small cell lung cancer on unenhanced computed tomography: Initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging* **10** 137–143.
- GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. [MR2596440](#)
- GUTMAN, D. A., COOPER, L. A., HWANG, S. N., HOLDER, C. A., GAO, J., AURORA, T. D., DUNN JR., W. D., SCARPACE, L., MIKKELSEN, T. et al. (2013). MR imaging predictors of molecular profile and survival: Multi-institutional study of the TCGA glioblastoma data set. *Radiology* **267** 560–569.
- HARALICK, R. M. (1979). Statistical and structural approaches to texture. *Proc. IEEE* **67** 786–804.
- HARALICK, R. M., SHANMUGAM, K. and DINSTEN, I. H. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern. Syst.* **6** 610–621.
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. [MR1082147](#)
- HODGSON, J. G., YEH, R.-F., RAY, A., WANG, N. J., SMIRNOV, I., YU, M., HARIONO, S., SILBER, J., FEILER, H. S. et al. (2009). Comparative analyses of gene copy number and mRNA expression in glioblastoma multiforme tumors and xenografts. *Neuro-Oncol.* **11** 477–487.
- JENNINGS, E. M., MORRIS, J. S., MANYAM, G. C., CARROLL, R. J. and BALADANDAYUTHAPANI, V. (2015). Bayesian models for flexible integrative analysis of multi-platform genomics data. In *Integrating Omics Data: Statistical and Computational Methods*, 1st ed. Cambridge University Press, Cambridge.
- HU, L. S., NING, S., ESCHBACHER, J. M., BAXTER, L. C., GAW, N., RANJBAR, S., PLASENCIA, J., DUECK, A. C., PENG, S. et al. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro-Oncol.* **19** 128–137.

- JENNINGS, E. M., MORRIS, J. S., CARROLL, R. J., MANYAM, G. C. and BALADANDAYUTHAPANI, V. (2012). Hierarchical Bayesian methods for integration of various types of genomics data. In *Genomic Signal Processing and Statistics (GENSIPS)*, 2012 IEEE International Workshop on 5–8. IEEE, New York.
- KICKINGEREDER, P., BURTH, S., WICK, A., GÖTZ, M., EIDEL, O., SCHLEMMER, H.-P., MAIER-HEIN, K. H., WICK, W., BENDSZUS, M. et al. (2016). Radiomic profiling of glioblastoma: Identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology* **280** 880–889.
- KIM, Y.-H., LACHUER, J., MITTELBRONN, M., PAULUS, W., BROKINKEL, B., KEYVANI, K., SURE, U., WREDE, K., NOBUSAWA, S. et al. (2011). Alterations in the RB1 pathway in low-grade diffuse gliomas lacking common genetic alterations. *Brain Pathol.* **21** 645–651.
- LANCKRIET, G. R., DE BIE, T., CRISTIANINI, N., JORDAN, M. I. and NOBLE, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics* **20** 2626–2635.
- LEE, J., JAIN, R., KHALIL, K., GRIFFITH, B., BOSCA, R., RAO, G. and RAO, A. (2016). Texture feature ratios from relative CBV maps of perfusion MRI are associated with patient survival in glioblastoma. *Am. J. Neuroradiol.* **37** 37–43.
- MANYAM, G., IVAN, C., CALIN, G. A. and COOMBES, K. R. (2013). targetHub: A programmable interface for miRNA–gene interactions. *Bioinformatics* **29** 2657–2658.
- MCGUFFEY, E., MORRIS, J. S., CARROLL, R. J., MANYAM, G. and BALADANDAYUTHAPANI, V. (2018). piBAG: Pathway-based integrative Bayesian modeling of multiplatform genomics data. Technical report.
- MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. [MR0997578](#)
- MONTANO, N., CENCI, T., MARTINI, M., D’ALESSANDRIS, Q. G., PELACCHI, F., RICCI-VITIANI, L., MAIRA, G., MARIA, R. D., LAROCCA, L. M. et al. (2011). Expression of EGFRvIII in glioblastoma: Prognostic significance revisited. *Neoplasia* **13** 1113–1121.
- NAKAMURA, M., KONISHI, N., TSUNODA, S., HIASA, Y., TSUZUKI, T., INUI, T. and SAKAKI, T. (1997). Retinoblastoma protein expression and MIB-1 correlate with survival of patients with malignant astrocytoma. *Cancer* **80** 242–249.
- NAKAMURA, M., YONEKAWA, Y., KLEIHUES, P. and OHGAKI, H. (2001). Promoter hypermethylation of the RB1 gene in glioblastomas. *Lab. Invest.* **81** 77–82.
- NAZARENKO, I., HEDE, S.-M., HE, X., HEDRÉN, A., THOMPSON, J., LINDSTRÖM, M. S. and NISTÉR, M. (2012). PDGF and PDGF receptors in glioma. *Ups. J. Med. Sci.* **117** 99–112.
- NETWORK, C. G. A. T. R. et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061.
- NICOLASJILWAN, M., HU, Y., YAN, C., MEERZAMAN, D., HOLDER, C. A., GUTMAN, D., JAIN, R., COLEN, R., RUBIN, D. L. et al. (2015). Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. *J. Neuroradiol.* **42** 212–221.
- OLIVARES, R. J., RAO, A., RAO, G., MORRIS, J. S. and BALADANDAYUTHAPANI, V. (2013). Integrative analysis of multi-modal correlated imaging-genomics data in glioblastoma. In *Genomic Signal Processing and Statistics (GENSIPS)*, 2013 IEEE International Workshop on 5–8. IEEE, New York.
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#)
- RU, Y., KECHRIS, K. J., TABAKOFF, B., HOFFMAN, P., RADCLIFFE, R. A., BOWLER, R., MAHAFFEY, S., ROSSI, S., CALIN, G. A. et al. (2014). The multiMiR R package and database: Integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Res.* **42** e133–e133.

- SOLOMON, D. A., KIM, J.-S., JENKINS, S., RESSOM, H., HUANG, M., COPPA, N., MABANTA, L., BIGNER, D., YAN, H. et al. (2008). Identification of p18INK4c as a tumor suppressor gene in glioblastoma multiforme. *Cancer Res.* **68** 2564–2569.
- STINGO, F. C., GUINDANI, M., VANNUCCI, M. and CALHOUN, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *J. Amer. Statist. Assoc.* **108** 876–891. [MR3174670](#)
- STUPP, R., HEGI, M. E., MASON, W. P., VAN DEN BENT, M. J., TAPHOORN, M. J., JANZER, R. C., LUDWIN, S. K., ALLGEIER, A., FISHER, B. et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* **10** 459–466.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TOMCZAK, K., CZERWIŃSKA, P. and WIZNEROWICZ, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19** A68.
- VERHAAK, R. G., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17** 98–110.
- WANG, W., BALADANDAYUTHAPANI, V., MORRIS, J. S., BROOM, B. M., MANYAM, G. and DO, K.-A. (2013). IBAG: Integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29** 149–159.
- WEI, L. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Stat. Med.* **11** 1871–1879.
- WOOD, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. *R News* **1** 20–25.
- YIN, D., OGAWA, S., KAWAMATA, N., TUNICI, P., FINOCCHIARO, G., EOLI, M., RUCKERT, C., HUYNH, T., LIU, G. et al. (2009). High-resolution genomic copy number profiling of glioblastoma multiforme by single nucleotide polymorphism DNA microarray. *Mol. Cancer Res.* **7** 665–677.
- ZHANG, Y., MORRIS, J. S., AERRY, S. N., RAO, A. U. and BALADANDAYUTHAPANI, V. (2019). Supplement to “Radio-iBAG: Radiomics-based integrative Bayesian analysis of multiplatform genomic data.” DOI:[10.1214/19-AOAS1238SUPP](#).
- ZHOU, M., HALL, L., GOLDFOF, D., RUSSO, R., BALAGURUNATHAN, Y., GILLIES, R. and GATENBY, R. (2014). Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: Preliminary results. *Transl. Oncol.* **7** 5–13.
- ZHU, Y., QIU, P. and JI, Y. (2014). TCGA-assembler: Open-source software for retrieving and processing TCGA data. *Nat. Methods* **11** 599–600.
- ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. [MR2137327](#)
- ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. [MR2252527](#)

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

- ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. [MR1224394](#)
- BILLER, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *J. Comput. Graph. Statist.* **9** 122–140. [MR1819868](#)
- BILLER, C. and FAHRMEIR, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Stat. Model.* **1** 195–211.
- BREZGER, A. and LANG, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Comput. Statist. Data Anal.* **50** 967–991. [MR2210741](#)
- CENTERS FOR DISEASE CONTROL AND PREVENTION (2017). HIV and viral hepatitis. South Carolina State Documents Depository.
- CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34** 305–334. [MR0888070](#)
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.

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

- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#)
- CHIPMAN, H. and MCCULLOCH, R. (2010). BayesTree: Bayesian methods for tree based models. R package version 0.3-1.1. Available at <http://CRAN.R-project.org/package=BayesTree>.
- DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998a). Automatic Bayesian curve fitting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 333–350. [MR1616029](#)
- DENISON, D. G., MALLICK, B. K. and SMITH, A. F. (1998b). Bayesian mars. *Stat. Comput.* **8** 337–346.
- EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statist. Sci.* **11** 89–121. [MR1435485](#)
- FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67. [MR1091842](#)
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. [MR2027492](#)
- GREEN, D. P. and KERN, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees. *Public Opin. Q.* **76** 491–511.
- GÜNTARD, H. F., SAAG, M. S., BENSON, C. A., DEL RIO, C., ERON, J. J., GALLANT, J. E., HOY, J. F., MUGAVERO, M. J., SAX, P. E. et al. (2016). Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2016 recommendations of the International Antiviral Society—USA panel. *J. Amer. Medical Assoc.* **316** 191–210.
- HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2018). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Available at [arXiv:1706.09523](https://arxiv.org/abs/1706.09523).
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. CRC Press, London. [MR1082147](#)
- HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting. *Statist. Sci.* **15** 196–223. [MR1820768](#)
- HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. [MR2816546](#)
- HOLMES, C. C. and MALLICK, B. K. (2001). Bayesian regression with multivariate linear splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 3–17. [MR1811987](#)
- LO RE, V., ZELDOW, B., KALLAN, M. J., TATE, J. P., CARBONARI, D. M., HENNESSY, S., KOSTMAN, J. R., LIM, J. K., GOETZ, M. B. et al. (2017). Risk of liver decompensation with cumulative use of mitochondrial toxic nucleoside analogues in HIV/hepatitis C virus coinfection. *Pharmacoepidemiol. Drug Saf.* **26** 1172–1181.
- MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis. Springer Series in Statistics*. Springer, Cham. [MR3309338](#)
- NATIONAL INSTITUTES OF HEALTH (2018). Panel on antiretroviral guidelines for adults and adolescents. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents Living with HIV, Dept. Health and Human Services. Available at <http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf>. Accessed: 2019-03-01.
- RASMUSSEN, C. E. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. [MR1293185](#)
- ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN, 1997)*. *IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. [MR1731682](#)
- ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 550–560.
- SAARELA, O., BELZILE, L. R. and STEPHENS, D. A. (2016). A Bayesian view of doubly robust causal inference. *Biometrika* **103** 667–681. [MR3551791](#)

- SCOURFIELD, A., JACKSON, A., WATERS, L., GAZZARD, B. and NELSON, M. (2011). The value of screening HIV-infected individuals for didanosine-related liver disease? *Antivir. Ther.* **16** 941–942.
- SORIANO, V., PUOTI, M., GARCIA-GASCÓ, P., ROCKSTROH, J. K., BENHAMOU, Y., BARREIRO, P. and MCGOVERN, B. (2008). Antiretroviral drugs and liver injury. *AIDS* **22** 1–13.
- STERLING, R. K., LISSEN, E., CLUMECK, N., SOLA, R., CORREA, M. C., MONTANER, J., SULKOWSKI, M. S., TORRIANI, F. J., DIETERICH, D. T. et al. (2006). Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology* **43** 1317–1325.
- VANSTEELANDT, S. and GOETGHEBEUR, E. (2003). Causal inference with generalized structural mean models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 817–835. [MR2017872](#)
- VANSTEELANDT, S. and JOFFE, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statist. Sci.* **29** 707–731. [MR3300367](#)
- VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics.* Springer, New York. [MR2867111](#)
- WOOD, S. and WOOD, M. S. (2015). Package ‘mgcv.’ R package Version 1-7. Available at <http://CRAN.R-project.org/package=mgcv>.
- ZELDOW, B., LO RE III, V. and ROY, J. (2019). Supplement to “A semiparametric modeling approach using Bayesian Additive Regression Trees with an application to evaluate heterogeneous treatment effects.” DOI:[10.1214/19-AOAS1266SUPP](https://doi.org/10.1214/19-AOAS1266SUPP).

The Annals of Applied Statistics

Next Issues

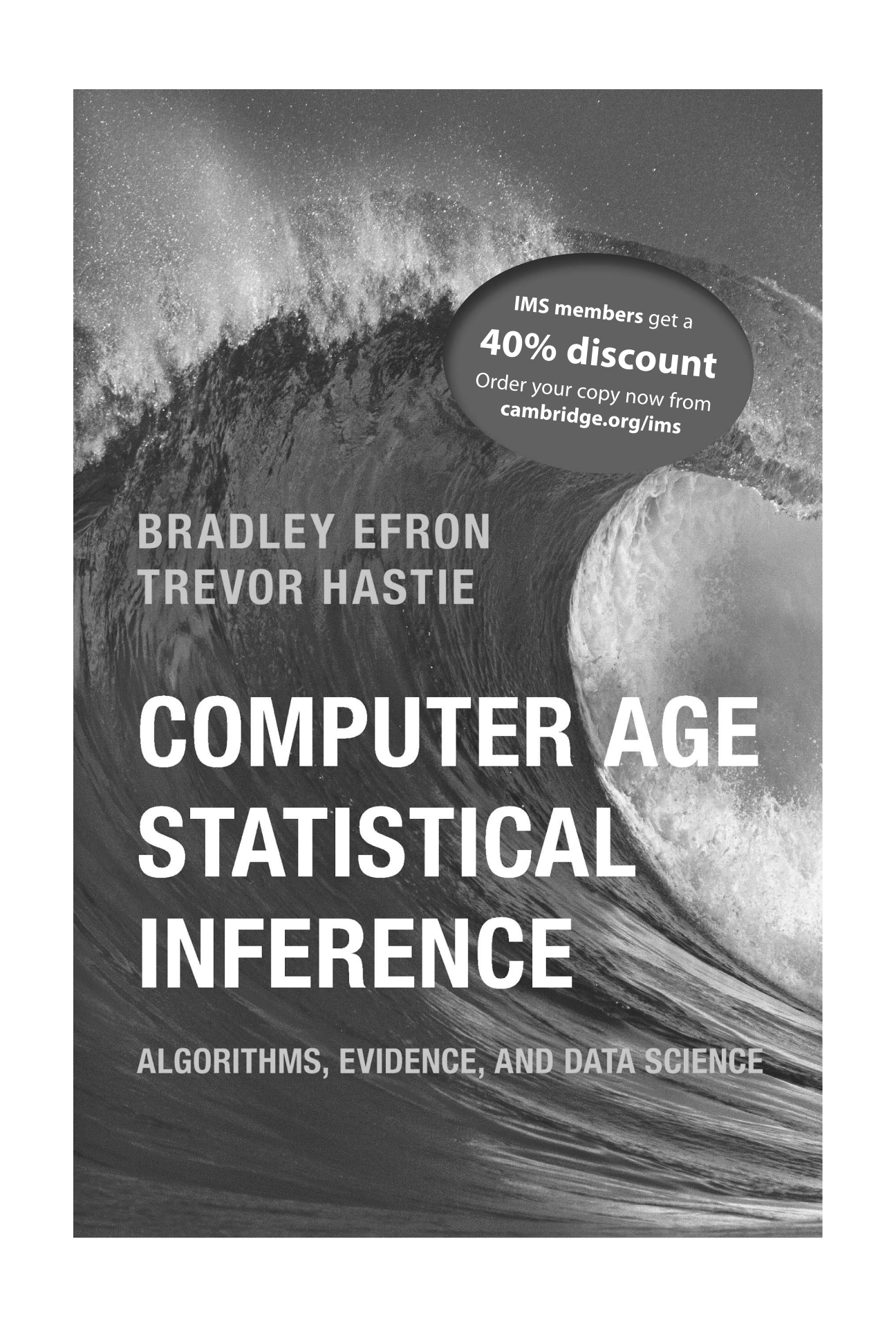
- Fitting a deeply-nested hierarchical model to a large book review dataset using a moment-based estimator NINGSHAN ZHANG, KYLE SCHMAUS AND PATRICK O. PERRY
- Bayesian factor models for probabilistic cause of death assessment with verbal autopsies TSUYOSHI KUNIHAMA, ZEHANG LI, SAMUEL CLARK AND TYLER 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 VITTEERT, 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 AMEIJERAS-ALONSO, AKLI BENALI, ROSA M. CRUJEIRAS,
ALBERTO RODRÍGUEZ-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?
PAUL 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
GIANLUCA 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



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE