ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Articles

continued

ISSN 1932-6157 (print)
ISSN 1941-7330 (online)

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 481–503
DOI: 10.1214/16-AOAS960

SPACE AND CIRCULAR TIME LOG GAUSSIAN COX PROCESSES WITH APPLICATION TO CRIME EVENT DATA

BY SHINICHIRO SHIROTA[1] AND ALAN E. GELFAND

Duke University

We view the locations and times of a collection of crime events as a space–time point pattern. Then, with either a nonhomogeneous Poisson process or with a more general Cox process, we need to specify a space–time intensity. For the latter, we need a *random* intensity which we model as a realization of a spatio-temporal log Gaussian process. Importantly, we view time as circular not linear, necessitating valid separable and nonseparable covariance functions over a bounded spatial region crossed with circular time. In addition, crimes are classified by crime type. Furthermore, each crime event is recorded by day of the year, which we convert to day of the week marks.

The contribution here is to develop models to accommodate such data. Our specifications take the form of hierarchical models which we fit within a Bayesian framework. In this regard, we consider model comparison between the nonhomogeneous Poisson process and the log Gaussian Cox process. We also compare separable vs. nonseparable covariance specifications.

Our motivating dataset is a collection of crime events for the city of San Francisco during the year 2012. We have location, hour, day of the year, and crime type for each event. We investigate models to enhance our understanding of the set of incidences.

REFERENCES

ADAMS, R. P., MURRAY, I. and MACKAY, D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the* 26*th International Conference on Machine Learning* MIT Press, Cambridge, MA.

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

BRANTINGHAM, P. and BRANTINGHAM, P. (1995). Criminality of place: Crime generators and crime attractors. *Eur. J. Crim. Policy Res.* **3** 5–26.

BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 823–841. MR1872069

CHAINEY, S., TOMPSON, L. and UHLIG, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Secur. J.* **21** 4–28.

CRESSIE, N. and HUANG, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. MR1731494

Key words and phrases. Derived covariates, hierarchical model, marked point pattern, Markov chain Monte Carlo, separable and nonseparable covariance functions, wrapped circular variables.

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261.

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, 2nd ed. Springer, New York. MR1950431

DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II: General Theory and Structure*, 2nd ed. Springer, Berlin. MR2371524

DOORNIK, J. (2007). *Ox: Object Oriented Matrix Programming*. Timberlake Consultants Press.

FISHER, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge Univ. Press, Cambridge. MR1251957

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 123–214. MR2814492

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space–time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475

GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice*. Wiley, Chichester. MR2384630

JAMMALAMADAKA, S. R. and SENGUPTA, A. (2001). *Topics in Circular Statistics*. World Scientific, River Edge, NJ. MR1836122

JONA-LASINIO, G., GELFAND, A. E. and JONA-LASINIO, M. (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *Ann. Appl. Stat.* **6** 1478–1498. MR3058672

LEININGER, T. J. (2014). Bayesian analysis of spatial point patterns. PhD dissertation. MR3232305

LEININGER, T. J. and GELFAND, A. E. (2016). Bayesian inference and model assessment for spatial point patterns using posterior predictive samples. *Bayesian Anal.* **12** 1–30. MR3597565

LIANG, W. W. J., COLVIN, J. B., SANSÓ, B. and LEE, H. K. H. (2014). Modeling and anomalous cluster detection for point processes using process convolutions. *J. Comput. Graph. Statist.* **23** 129–150. MR3173764

MARDIA, K. V. (1972). *Statistics of Directional Data*. Academic Press, London. MR0336854

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. Wiley, Chichester. MR1828667

MOHLER, G. O. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1825–1839. MR3127957

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019

MURRAY, I. and ADAMS, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems* 23. MIT Press, Cambridge, MA.

MURRAY, I., ADAMS, R. P. and GRAHAM, M. M. (2010). Elliptical slice sampling. In *Proceedings of the* 13*th International Conference on Artifical Intelligence and Statistics* (*AISTAT*). AISTAT Press.

OGATA, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.

PORCU, E., BEVILACQUA, M. and GENTON, M. G. (2016). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *J. Amer. Statist. Assoc.* **111** 888–898. MR3538713

RODRIGUES, A. and DIGGLE, P. J. (2012). Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *J. Amer. Statist. Assoc.* **107** 93–101. MR2949344

SHIROTA, S. and GELFAND, A. E. (2017). Supplement to "Space and circular time log Gaussian Cox processes with application to crime event data." DOI:10.1214/16-AOAS960SUPP.

TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc.* **105** 1403–1417. MR2796559

WANG, F. and GELFAND, A. E. (2014). Modeling space and space–time directional data using projected Gaussian processes. *J. Amer. Statist. Assoc.* **109** 1565–1580.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 504–526
DOI: 10.1214/16-AOAS974

RANDOM EFFECTS MODELS FOR IDENTIFYING THE MOST HARMFUL MEDICATION ERRORS IN A LARGE, VOLUNTARY REPORTING DATABASE

BY SERGIO VENTURINI*, JESSICA M. FRANKLIN†,‡, LAURA MORLOCK§ AND FRANCESCA DOMINICI¶,[1]

*Università Commerciale Luigi Bocconi**, *Brigham and Women's Hospital*†, *Harvard Medical School*‡, *Johns Hopkins University*§ *and Harvard TH Chan School of Public Health*¶

Medical errors are a major source of preventable morbidity, mortality and healthcare costs. Voluntary reporting systems are useful data sources that collect detailed information on the circumstances of medical errors occurring in hospitals. Identifying the characteristics of errors that frequently result in patient harm when they occur would allow investigators to prioritize among the many sources of potential errors and design targeted prevention strategies. In this paper, we use data from MEDMARX, a large anonymous and voluntary reporting system for medication errors, to identify the combinations of error characteristics that are more likely to result in harm. To this end, we consider a Bayesian hierarchical model with crossed random effects and a flexible specification of the random effects distribution. We then provide a ranking of the errors using optimal Bayesian ranking based on their probability of harm. The use of optimal Bayesian ranking accounts for the varying amount of uncertainty across the random effects estimates. Finally, we examine the sensitivity of results to different specifications of the random effects distributions. The utility of flexible random effects assumptions is illustrated by empirically comparing results under several choices. We found that errors caused by mistakes in reconciling a patient's current medication list with the medications prescribed at hospital discharge have an estimated 10.5% probability of harm. These errors had the highest rate of harm of errors that occur during the prescribing stage of medication use. In addition, we found that the results are sensitive to the random effects distribution used in estimation. Thus, an approach that explores this sensitivity is important for accurately comparing the relative harm across errors.

REFERENCES

ABAD, A. A., LITIÈRE, S. and MOLENBERGHS, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostat*. **11** 771–786.

AHMED, I., BÉGAUD, B. and TUBERT-BITTER, P. (2015). Evaluation of post-marketing safety using spontaneous reporting databases. In *Statistical Methods for Evaluating Safety in Medical Product Development* (A. L. Gould, ed.). Wiley, New York.

Key words and phrases. Bayesian hierarchical model, empirical Bayes, data mining, spontaneous reporting.

AHMED, I., HARAMBURU, F., FOURRIER-RÉGLAT, A., THIESSARD, F., KREFT-JAIS, C., MIREMONT-SALAMÉ, G., BÉGAUD, B. and TUBERT-BITTER, P. (2009). Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. *Stat. Med.* **28** 1774–1792. MR2751597

AHMED, I., DALMASSO, C., HARAMBURU, F., THIESSARD, F., BRÖET, P. and TUBERT-BITTER, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* **66** 301–309.

AHMED, I., THIESSARD, F., MIREMONT-SALAMÉ, G., HARAMBURU, F., KREFT-JAIS, C., BÉGAUD, B. and TUBERT-BITTER, P. (2012). Early detection of pharmacovigilance signals with automated methods based on false discovery rates. *Drug Saf.* **35** 495–506.

ASPDEN, P., CORRIGAN, J. M., WOLCOTT, J. and ERICKSON, S. M. (2003). *Patient Safety: Achieving a New Standard for Care*. The National Academy Press, Washington, DC.

BAYARRI, M. J. and CASTELLANOS, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statist. Sci.* **22** 322–343. MR2416808

BOOCKVAR, K. S., CARLSON LACORTE, H., GIAMBANCO, V., FRIDMAN, B. and SIU, A. (2006). Medication reconciliation for reducing drug-discrepancy adverse events. *Am. J. Geriatr. Pharmacother.* **4** 236–243.

BRENNAN, T., LEAPE, L., LAIRD, N., HEBERT, L., LOCALIO, A., LAWTHERS, A., NEWHOUSE, J., WEILER, P. and HIATT, H. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard medical practice study I. *N. Engl. J. Med.* **324** 370–376.

BROWNE, W. J. (2004). An illustration of the use of reparameterisation methods for improving MCMC efficiency in crossed random effect models. *Multilevel Model. Newsl.* **16** 13–25.

DEY, D. K., GELFAND, A. E., SWARTZ, T. B. and VLACHOS, P. K. (1998). A simulation-intensive approach for checking hierarchical models. *TEST* **7** 325–346.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99** 205–215. MR2054299

DUMOUCHEL, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist.* **53** 177–190.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

EUGÉNE, V. P., EGBERTS, A., HEERDINK, E. R. and LEUFKENS, H. (2000). Detecting drug–drug interactions using a database for spontaneous adverse drug reactions: An example with diuretics and non-steroidal anti-inflammatory drugs. *Eur. J. Clin. Pharmacol.* **56** 733–738.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FERNÁNDEZ, C. and STEEL, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93** 359–371. MR1614601

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure. *Statist. Sci.* **23** 250–260. MR2516823

GBD 2013 MORTALITY AND CAUSES OF DEATH COLLABORATORS (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **385** 117–171.

GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, Cambridge.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., GOEGEBEUR, Y., TUERLINCKX, F. and VAN MECHELEN, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 247–268. MR1821324

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

GIBBONS, R. D. and AMATYA, A. K. (2016). *Statistical Methods for Drug Safety*. CRC Press, Boca Raton, FL.

GIBBONS, R. D., SEGAWA, E., KARABATSOS, G., AMATYA, A. K., BHAUMIK, D. K., BROWN, C. H., KAPUR, K., MARCUS, S. M., HUR, K. and MANN, J. J. (2008). Mixed-effects Poisson regression analysis of adverse event reports: The relationship between antidepressants and suicide. *Stat. Med.* **27** 1814–1833. MR2420347

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HUANG, X. (2011). Detecting random-effects model misspecification via coarsened data. *Comput. Statist. Data Anal.* **55** 703–714. MR2736589

HUCKELS-BAUMGART, S. and MANSER, T. (2014). Identifying medication error chains from critical incident reports: A new analytic approach. *J. Clin. Pharmacol.* **54** 1188–1197.

KOHN, L. T., CORRIGAN, J. and DONALDSON, M. S. (2000). *To Err Is Human*: *Building a Safer Health System*. The National Academy Press, Washington, DC.

THE JOINT COMMISSION (2000). Reporting of Medical/Health Care Errors: A Position Statement of the Joint Commission.

KWAN, J. L., LO, L., SAMPSON, M. and SHOJANIA, K. G. (2013). Medication reconciliation during transitions of care as a patient safety strategy: A systematic review. *Ann. Intern. Med.* **158** 397–403.

KYUNG, M., GILL, J. and CASELLA, G. (2010). Estimation in Dirichlet random effects models. *Ann. Statist.* **38** 979–1009. MR2604702

LEAPE, L., BRENNAN, T., LAIRD, N., LAWTHERS, A., LOCALIO, A., BARNES, B., HEBERT, L., NEWHOUSE, J., WEILER, P. and HIATT, H. (1991). The nature of adverse events in hospitalized patients. Results of the Harvard medical practice study II. *N. Engl. J. Med.* **324** 377–384.

LEE, K. J. and THOMPSON, S. G. (2008). Flexible parametric models for random-effects distributions. *Stat. Med.* **27** 418–434. MR2418453

LOUIS, T. A. and SHEN, W. (1999). Innovations in Bayes and empirical Bayes methods: Estimating parameters, populations and ranks. *Stat. Med.* **18** 2493–2505.

MACEACHERN, S. N. and MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–238.

MACEACHERN, S. N. and PERUGGIA, M. (2000). Importance link function estimation for Markov chain Monte Carlo methods. *J. Comput. Graph. Statist.* **9** 99–121. MR1819867

MADIGAN, D., RYAN, P., SIMPSON, S. and ZORYCH, I. (2010). Bayesian methods in pharmacovigilance. In *Bayesian Statistics*, *Vol.* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford Univ. Press, London.

MORLOCK, L., DOMINICI, F., MYERS, J. A., SHORE, A. D., PRONOVOST, P. J., DY, S. M. and COUSINS, D. D. (2010). Comparing near miss and harmful medication errors: Testing the causal continuum hypothesis using data from the MEDMARX National Reporting System, Technical report, Johns Hopkins Univ., Baltimore, MD.

MUELLER, S. K., SPONSLER, K. C., KRIPALANI, S. and SCHNIPPER, J. L. (2012). Hospital-based medication reconciliation practices: A systematic review. *Arch. Intern. Med.* **172** 1057–1069.

MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, Cham. MR3309338

NATIONAL COORDINATING COUNCIL FOR MEDICATION ERROR REPORTING AND PREVENTION (NCC MERP) (2001). *Medication Error Index*.

PRONOVOST, P., WEAST, B., SCHWARZ, M., WYSKIEL, R. M., PROW, D., MILANOVICH, S. N., BERENHOLTZ, S., DORMAN, T. and LIPSETT, P. (2003). Medication reconciliation: A practical tool to reduce the risk of medication errors. *J. Crit. Care* **18** 201–205.

RAMJAUN, A., SUDARSHAN, M., PATAKFALVI, L., TAMBLYN, R. and MEGUERDITCHIAN, A. N. (2015). Educating medical trainees on medication reconciliation: A systematic review. *BMC Med. Educ.* **15** 33.

SANTELL, J. P., HICKS, R. W., MCMEEKIN, J. and COUSINS, D. D. (2003). Medication errors: Experience of the United States Pharmacopeia (USP) MEDMARX Reporting System. *J. Clin. Pharmacol.* **43** 760–767.

SCHIFF, G. D., AMATO, M. G., EGUALE, T., BOEHNE, J. J., WRIGHT, A., KOPPEL, R., RASHIDEE, A. H., ELSON, R. B., WHITNEY, D. L., THACH, T.-T., BATES, D. W. and SEGER, A. C. (2015). Computerised physician order entry-related medication errors: Analysis of reported errors and vulnerability testing of current systems. *BMJ Qual. Saf.* **24** 264–271.

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061

SINHARAY, S. and STERN, H. S. (2003). Posterior predictive model checking in hierarchical models. *J. Statist. Plann. Inference* **111** 209–221. MR1955882

STERN, H. S. and CRESSIE, N. (2000). Posterior predictive model checks for disease mapping models. *Stat. Med.* **19** 2377–2397.

TCHETGEN, E. J. and COULL, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika* **93** 1003–1010. MR2285086

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017a). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPA.

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017b). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPB.

VENTURINI, S., FRANKLIN, J. M., MORLOCK, L. and DOMINICI, F. (2017c). Supplement to "Random effects models for identifying the most harmful medication errors in a large, voluntary reporting database." DOI:10.1214/16-AOAS974SUPPC.

WAAGEPETERSEN, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scand. J. Stat.* **33** 721–731. MR2300912

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 527–552
DOI: 10.1214/16-AOAS987

ASSESSING SYSTEMATIC RISK IN THE S&P500 INDEX BETWEEN 2000 AND 2011: A BAYESIAN NONPARAMETRIC APPROACH[1]

BY ABEL RODRÍGUEZ*, ZIWEI WANG† AND ATHANASIOS KOTTAS*

University of California, Santa Cruz and IAC Publishing Labs†*

We develop a Bayesian nonparametric model to assess the effect of systematic risks on multiple financial markets, and apply it to understand the behavior of the S&P500 sector indexes between January 1, 2000 and December 31, 2011. More than prediction, our main goal is to understand the evolution of systematic and idiosyncratic risks in the U.S. economy over this particular time period, leading to novel sector-specific risk indexes. To accomplish this goal, we model the appearance of extreme losses in each market using a superposition of two Poisson processes, one that corresponds to systematic risks that are shared by all sectors, and one that corresponds to the idiosyncratic risk associated with a specific sector. In order to capture changes in the risk structure over time, the intensity functions associated with each of the underlying components are modeled using a Dirichlet process mixture model. Among other interesting results, our analysis of the S&P500 index suggests that there are few idiosyncratic risks associated with the consumer staples sector, whose extreme negative log returns appear to be driven mostly by systematic risks.

REFERENCES

ACHARYA, V., ENGLE, R. and RICHARDSON, M. (2012). Capital shortfall: A new approach to ranking and regulating systemic risks. *Am. Econ. Rev.* **102** 59–64.

ACHARYA, V. V., PEDERSEN, L. H., PHILIPPON, T. and RICHARDSON, M. P. (2010). Measuring systemic risk. Technical report, AFA 2011 Denver Meetings Paper.

ADRIAN, T. and BRUNNERMEIER, M. (2010). Covar: A systemic risk contribution measure. Technical report, Princeton Univ., Princeton, NJ.

ALLEN, D., GERRANS, P., SINGH, A. and POWELL, R. (2009). Quantile regression: Its application in investment analysis. *Journal of the Securities Institute of Australia* **1** 1–12.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR365969

BARNES, M. L. and HUGHES, W. A. (2002). A quantile regression analysis of the cross section of stock market returns. Technical report, Federal Reserve Bank of Boston.

BASSETTI, F., CASARIN, R. and LEISEN, F. (2014). Beta-product dependent Pitman–Yor processes for Bayesian inference. *J. Econometrics* **180** 49–72. MR3188911

CHANG, M. C., HUNG, J. C. and NIEH, C. C. (2011). Reexamination of capital asset pricing model (CAPM): An application of quantile regression. *African Journal of Business Management* **5** 12684–12690.

Key words and phrases. Dirichlet process mixture modeling, nonhomogeneous Poisson process, nonparametric Bayes, systematic risk.

COX, D. R. (1955). Some statistical methods connected with series of events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **17** 129–157; discussion, 157–164. MR0092301

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd ed. Springer, New York. MR1950431

DANIEL, K. D., HIRSHLEIFER, D. and SUBRAHMANYAM, A. (2001). Overconfidence, arbitrage, and equilibrium asset pricing. *J. Finance* **56** 921–965.

DIACONIS, P. and YLVISAKER, D. (1985). Quantifying prior opinion. In *Bayesian Statistics* 2 (J. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 133–156. North-Holland, Amsterdam. MR0862488

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FAMA, E. F. and FRENCH, K. R. (1992). The cross-section of expected stock returns. *J. Finance* **47** 427–465.

FAMA, E. F. and FRENCH, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives* **18** 25–46.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FRENCH, C. W. (2003). The Treynor capital asset pricing model. *Journal of Investment Management* **1** 60–72.

GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028

GELMAN, A. and RUBIN, D. B. (1992). Inferences from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GRIFFIN, J. E. and WALKER, S. G. (2011). Posterior simulation of normalized random measure mixtures. *J. Comput. Graph. Statist.* **20** 241–259.

GRIFFITHS, R. C. and MILNE, R. K. (1978). A class of bivariate Poisson processes. *J. Multivariate Anal.* **8** 380–395. MR0512608

HATJISPYROS, S. J., NICOLERIS, T. and WALKER, S. G. (2011). Dependent mixtures of Dirichlet processes. *Comput. Statist. Data Anal.* **55** 2011–2025. MR2785111

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

ISHWARAN, H. and ZAREPOUR, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika* **87** 371–390. MR1782485

KALLI, M., GRIFFIN, J. E. and WALKER, S. G. (2011). Slice sampling mixture models. *Stat. Comput.* **21** 93–105. MR2746606

KOLOSSIATIS, M., GRIFFIN, J. E. and STEEL, M. F. J. (2013). On Bayesian nonparametric modelling of two correlated distributions. *Stat. Comput.* **23** 1–15. MR3018346

KOTTAS, A. and SANSÓ, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference* **137** 3151–3163. MR2365118

KOTTAS, A., WANG, Z. and RODRÍGUEZ, A. (2012). Spatial modeling for risk assessment of extreme values from environmental time series: A Bayesian nonparametric approach. *Environmetrics* **23** 649–662. MR3019057

KOTTAS, A., BEHSETA, S., MOORMAN, D. E., POYNOR, V. and OLSON, C. R. (2012). Bayesian nonparametric analysis of neuronal intensity rates. *Journal of Neuroscience Methods* **203** 241–253.

LANDO, D. (1998). On Cox processes and credit risk securities. *Review of Derivatives Research* **2** 99–120.

LEISEN, F. and LIJOI, A. (2011). Vectors of two-parameter Poisson–Dirichlet processes. *J. Multivariate Anal*. **102** 482–495. MR2755010

LEISEN, F., LIJOI, A. and SPANÓ, D. (2013). A vector of Dirichlet processes. *Electron. J. Stat*. **7** 62–90.

LIJOI, A. and NIPOTI, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *J. Amer. Statist. Assoc*. **109** 802–814. MR3223751

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014b). Dependent mixture models: Clustering and borrowing information. *Comput. Statist. Data Anal*. **71** 417–433. MR3131980

LIJOI, A. and PRÜNSTER, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) 80–136. Cambridge Univ. Press, Cambridge.

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical report, Dept. Statistics, Ohio State Univ., Columbus, OH.

MANDELBROT, B. (1963). The variation of certain speculative prices. *Journal of Business* **36** 394–419.

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **66** 735–749. MR2088779

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

RODRÍGUEZ, A., WANG, Z. and KOTTAS, A. (2017). Supplement to "Assessing systematic risk in the S& P500 index between 2000 and 2011: A Bayesian nonparametric approach." DOI:10.1214/16-AOAS987SUPP.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650.

TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist. Assoc*. **105** 1403–1417. MR2796559

TADDY, M. A. and KOTTAS, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Anal*. **7** 335–362.

TREYNOR, J. L. (1961). Market Value, Time, and Risk. Unpublished manuscript.

TREYNOR, J. L. (1962). Toward a Theory of Market Value of Risky Assets. Unpublished manuscript.

XIAO, S., KOTTAS, A. and SANSÓ, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *Ann. Appl. Stat*. **9** 353–382. MR3341119

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 553–597
DOI: 10.1214/16-AOAS993

A CONTINUOUS-TIME STOCHASTIC BLOCK MODEL FOR BASKETBALL NETWORKS[1]

BY LU XIN, MU ZHU AND HUGH CHIPMAN

Royal Bank of Canada, University of Waterloo and Acadia University

For professional basketball, finding valuable and suitable players is the key to building a winning team. To deal with such challenges, basketball managers, scouts and coaches are increasingly turning to analytics. Objective evaluation of players and teams has always been the top goal of basketball analytics. Typical statistical analytics mainly focuses on the box score and has developed various metrics. In spite of the more and more advanced methods, metrics built upon box score statistics provide limited information about how players interact with each other. Two players with similar box scores may deliver distinct team plays. Thus professional basketball scouts have to watch real games to evaluate players. Live scouting is effective, but suffers from inefficiency and subjectivity. In this paper, we go beyond the static box score and model basketball games as dynamic networks. The proposed continuous-time stochastic block model clusters the players according to their playing style and performance. The model provides cluster-specific estimates of the effectiveness of players at scoring, rebounding, stealing, etc., and also captures player interaction patterns within and between clusters. By clustering similar players together, the model can help basketball scouts to narrow down the search space. Moreover, the model is able to reveal the subtle differences in the offensive strategies of different teams. An application to NBA basketball games illustrates the performance of the model.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D Stat.* **24** 179–195.

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.

CERVONE, D., D'AMOUR, A., BORNN, L. and GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *J. Amer. Statist. Assoc.* **111** 585–599. MR3538688

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York.

DUBOIS, C., BUTTS, C. T. and SMYTH, P. (2013). Stochastic blockmodeling of relational event dynamics. In *Proceedings of the* 16*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*).

Key words and phrases. Clustering, transactional network, Markov chain, EM algorithm, Gibbs sampling, basketball analytics, social network.

FEWELL, J. H., ARMBRUSTER, D., INGRAHAM, J., PETERSEN, A. and WATERS, J. S. (2012). Basketball teams as strategic networks. *PLoS ONE* **7** 849–911.

HO, Q., SONG, L. and XING, E. P. (2011). Evolving cluster mixed-membership blockmodel for time-varying networks. In *Proceedings of the* 14*th International Conference on Artifical Intelligence and Statistics* (*AISTATS*).

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107. MR2788206

OLIVER, D. (2004). *Basketball on Paper*: *Rules and Tools for Performance Analysis*. Potomac Books, Inc., Dulles, VA.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SHAFIEI, M. and CHIPMAN, H. (2010). Mixed-membership stochastic block-models for transactional networks. In *Proceedings of the International Conference on Data Mining*.

SHEA, S. M. and BAKER, C. E. (2013). *Basketball Analytics*: *Objective and Efficient Strategies for Understanding How Teams Win*. Advanced Metrics, LLC, Lake St. Louis, MO.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

VU, D. Q., ASUNCION, A. U., HUNTER, D. R. and SMYTH, P. (2011). Continuous-time regression models for longitudinal networks. *Adv. Neural Inf. Process. Syst.*

WANG, Y. J. and WONG, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19. MR0883333

XU, K. S. and HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562.

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 598–637
DOI: 10.1214/16-AOAS996

ROBUST AND SCALABLE BAYESIAN ANALYSIS OF SPATIAL NEURAL TUNING FUNCTION DATA

BY KAMIAR RAHNAMA RAD[1],
TIMOTHY A. MACHADO[2] AND LIAM PANINSKI[3]

City University of New York, Cognescent Corporation and Columbia University

A common analytical problem in neuroscience is the interpretation of neural activity with respect to sensory input or behavioral output. This is typically achieved by regressing measured neural activity against known stimuli or behavioral variables to produce a "tuning function" for each neuron. Unfortunately, because this approach handles neurons individually, it cannot take advantage of simultaneous measurements from spatially adjacent neurons that often have similar tuning properties. On the other hand, sharing information between adjacent neurons can errantly degrade estimates of tuning functions across space if there are sharp discontinuities in tuning between nearby neurons. In this paper, we develop a computationally efficient block Gibbs sampler that effectively pools information between neurons to denoise tuning function estimates while simultaneously preserving sharp discontinuities that might exist in the organization of tuning across space. This method is fully Bayesian, and its computational cost per iteration scales subquadratically with total parameter dimensionality. We demonstrate the robustness and scalability of this approach by applying it to both real and synthetic datasets. In particular, an application to data from the spinal cord illustrates that the proposed methods can dramatically decrease the experimental time required to accurately estimate tuning functions.

REFERENCES

AFONSO, M. V., BIOUCAS-DIAS, J. M. and FIGUEIREDO, M. A. T. (2010). Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.* **19** 2345–2356. MR2798930

AHMADIAN, Y., PILLOW, J. W. and PANINSKI, L. (2011). Efficient Markov chain Monte Carlo methods for decoding neural spike trains. *Neural Comput.* **23** 46–96. MR2768274

AHRENS, M., ORGER, M., ROBSON, D., LI, J. and KELLER, P. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods* **10** 413–420.

AKAY, T., TOURTELLOTTE, W. G., ARBER, S. and JESSELL, T. M. (2014). Degradation of mouse locomotor pattern in the absence of proprioceptive sensory feedback. *Proc. Natl. Acad. Sci. USA* **111** 16877–16882.

ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. MR0359122

BARBERO, A. and SRA, S. (2011). Fast Newton-type methods for total variation regularization. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 313–320.

Key words and phrases. Computational neuroscience, scalability and robustness, Bayesian, posterior sampling, spatial statistics.

BARDSLEY, J. M., SOLONEN, A., HAARIO, H. and LAINE, M. (2014). Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. *SIAM J. Sci. Comput.* **36** A1895–A1910. MR3248038

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1993). Towards Bayesian image analysis. *J. Appl. Stat.* **20** 107–119.

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811

BOUCHARD, K. E., MESGARANI, N., JOHNSON, K. and CHANG, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495** 327–332.

BOUMAN, I. and LIU, B. (1988). A multiple resolution approach to regularization. In *Proceedings SPIE* 1001, *Visual Communications and Image Processing* '88 512–520.

BOUMAN, C. and SAUER, K. (1993). A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Process.* **2** 296–310.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

BRANDT, A. (1977). Multigrid Monte Carlo method. Conceptual foundations. *Math. Comp.* **31** 333–390.

BREZGER, A., FAHRMEIR, L. and HENNERFEIND, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *J. Roy. Statist. Soc. Ser. C* **56** 327–345. MR2370993

BUADES, A., COLL, B. and MOREL, J. M. (2005). A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.* **4** 490–530.

BUESING, L., MACKE, J. H. and SAHANI, M. (2012). Learning stable, regularised latent models of neural population dynamics. *Network* **23** 24–47.

CANDES, E. J. (1999a). Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.* **6** 197–218.

CANDES, E. J. (1999b). Curvelets—a surprisingly effective nonadaptive representation for objects with edges. In *Curve and Surface Fitting*: *Saint-Malo* (A. Cohen, C. Rabut and L. L. Schumaker, eds.) Vanderbilt Univ. Press, Nashville, TN.

CASELLA, G. (2001). Empirical Bayes Gibbs sampling. *Biostat.* **2** 485–500.

CASELLA, G., GHOSH, M., GILL, J. and KYUNG, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411.

CHAMBOLLE, A. (2004). An algorithm for total variation minimization and applications. *J. Math. Imaging Vision* **20** 89–97. MR2049783

CHARBONNIER, P., BLANC-FERAUD, L., AUBERT, G. and BARLAUD, M. (1997). Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.* **6** 298–311.

CRONIN, B., STEVENSON, I. H., SUR, M. and KORDING, P. (2010). Hierarchical Bayesian modeling and Markov Chain Monte Carlo sampling for tuning-curve analysis. *J. Neurophysiol.* **103** 591–602.

CUNNINGHAM, J., YU, B., SHENOY, K. and SAHANI, M. (2008). Inferring neural firing rates from spike trains using Gaussian processes. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds.). Curran Associates, Red Hook, NY.

CUNNINGHAM, J. P., GILJA, V., RYU, S. and SHENOY, K. (2009). Methods for estimating neural firing rates, and their application to brain-machine interface. *Neural Netw.* **22** 1235–1246.

CZANNER, G., EDEN, U., WIRTH, S., YANIKE, M., SUZUKI, W. and BROWN, E. (2008). Analysis of between-trial and within-trial neural spiking dynamics. *J. Neurophysiol.* **99** 2672–2693.

DABOV, K., FOI, A., KATKOVNIK, V. and EGIAZARIAN, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16** 2080–2095. MR2460626

DAVIS, T. (2006). *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA.

DAYAN, P. and ABBOTT, L. F. (2001). *Theoretical Neuroscience*. *Computational Neuroscience*. MIT Press, Cambridge, MA. MR1985615

DEFRISE, M., VANHOVE, C. and LIU, X. (2011). An algorithm for total variation regularization in high-dimensional linear problems. *Inverse Probl.* **27** 065002.

DOI, E., GAUTHIER, J. L., FIELD, G. D., SHLENS, J., SHER, A., GRESCHNER, M., MACHADO, T. A., JEPSON, L. H., MATHIESON, K., GUNNING, D. E., LITKE, A. M., PANINSKI, L., CHICHILNISKY, E. J. and SIMONCELLI, E. P. (2012). Efficient coding of spatial information in the primate retina. *J. Neurosci.* **32** 16256–16264.

DONOHO, D. and JOHNSTONE, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DUANE, S., KENNEDY, A. D., PENDELTON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **55**.

ELTOFT, T., KIM, T. and LEE, T. (2006). On the multivariate Laplace distribution. *IEEE Signal Process. Lett.* **13** 300–303.

FAHRMEIR, L., KNEIB, T. and LANG, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statist. Sinica* **14** 731–761. MR2087971

FEINBERG, E. H. and MEISTER, M. (2014). Orientation columns in the mouse superior colliculus. *Nature* **519** 229–232.

GAO, Y., BLACK, M., BIENENSTOCK, E., SHOHAM, S. and DONOGHUE, J. (2002). Probabilistic inference of arm motion from neural activity in motor cortex. In *Advances in Neural Information Processing Systems* 14 (Z. G. Thomas, G. Dietterich and S. Becker, eds.) 213–220.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (2003). *Bayesian Data Analysis*. CRC Press, Boca Raton, FL.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GEMAN, D. and REYNOLDS, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.* **14** 367–383.

GEMAN, D. and YANG, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.* **4** 932–946.

GEORGOPOULOS, A., KETTNER, R. and SCHWARTZ, A. (1986). Neuronal population coding of movement direction. *Science* **233** 1416–1419.

GILAVERT, C., MOUSSAOUI, S. and IDIER, J. (2015). Efficient Gaussian sampling for solving large-scale inverse problems using MCMC. *IEEE Trans. Signal Process.* **63** 70–80. MR3286325

GIRMAN, S. V., SAUVÉ, Y. and LUND, R. D. (1999). Receptive field properties of single neurons in rat primary visual cortex. *J. Neurophysiol.* **82** 301–311.

GIROLAMI, M., CALDERHEAD, B. and CHIN, S. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **73** 123–214. MR2814492

GOODMAN, J. and SOKAL, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D* **40** 2035–2071.

GREEN, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imag.* **9** 84–94.

GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **73** 304–321.

GROVES, A. R., CHAPPELL, M. A. and WOOLRICH, M. W. (2009). Combined spatial and non-spatial prior for inference on MRI time-scales. *NeuroImage* **45** 795–809.

GUILLAUME, F. and PENNY, W. D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage* **34** 1108–1125.

HAFTING, T., FYHN, M., MOLDEN, S., MOSER, M. B. and MOSER, E. I. (2005). Microstructure of a spatial map in the enthorhinal cortex. *Nature* **436** 801–806.

HALLAC, D., LESKOVEC, J. and BOYD, S. (2015). Network lasso: Clustering and optimization in large graphs. In *SIGKDD* 387–396.

HAMEL, E. J. O., GREWE, B. F., PARKER, J. G. and SCHNITZER, M. J. (2015). Cellular level brain imaging in behaving mammals: An engineering approach. *Neuron* **86** 140–159. DOI:10.1016/j.neuron.2015.03.055.

HARRISON, L. M. and GREEN, G. G. R. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50** 1126–1141.

HARRISON, L. M., PENNY, W., ASHBURNER, J., TRUJILLO-BARRETO, N. and FRISTON, K. J. (2007). Diffusion-based spatial priors for imaging. *NeuroImage* **38** 677–695.

HARRISON, S. J., WOOLRICH, M. W., ROBINSON, E. C., GLASSER, M. F., BECKMAN, C. F., JENKINSON, M. and SMITH, S. M. (2015). Large-scale probabilistic functional modes from resting state fMRI. *NeuroImage* **109** 217–231.

HENRY, G. H., DREHER, B. and BISHOP, P. O. (1974). Orientation specificity of cells in cat striate cortex. *J. Neurophysiol.* **37** 1394–1409.

HOFFMAN, Y. (2009). Gaussian fields and constrained simulations of the large-scale structure. In *Data Analysis in Cosmology* 565–583. Springer, Berlin.

HOFFMAN, Y. and RIBAK, E. (1991). Constrained realization of Gaussian fields—a simple algorithm. *Astrophys. J.* **380** L5–L8.

HUBEL, D. H. and WIESEL, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160** 106–154.

HUBEL, D. H. and WIESEL, T. N. (1968). Receptive fields and functional architecture of the monkey striate cortex. *J. Neurophysiol.* **195** 215–243.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415

ISSA, J. B., HAEFFELE, B. D., AGARWAL, A., BERGLES, D. E., YOUNG, E. D. and YUE, D. T. (2014). Multiscale optical Ca 2+ imaging of tonal organization in mouse auditory cortex. *Neuron* **83** 944–959.

KASCHUBE, M., SCHNABEL, M., LÖWEL, S., COPPOLA, D. M., WHITE, L. E. and WOLF, F. (2010). Universality in the evolution of orientation columns in the visual cortex. *Science* **330** 1113–1116.

KEIL, W., KASCHUBE, M., SCHNABEL, M., KISVARDAY, Z., LOWEL, S., COPPOLA, D. M., WHITE, L. E. and WOLF, F. (2012). Response to comment on "Universality in the evolution of orientation columns in the visual cortex." *Science* **336**.

KIM, S. J., KOH, S., BOYD, S. and GORINEVSKY, D. (2009). ℓ_1 trend filtering. *SIAM Rev.* **51** 339–360.

KROUCHEV, N., KALASKA, J. F. and DREW, T. (2006). Sequential activation of muscle synergies during locomotion in the intact cat as revealed by cluster analysis and direct decomposition. *J. Neurophysiol.* **96** 1991–2010.

LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877

LANG, S., FRONK, E.-M. and FAHRMEIR, L. (2002). Function estimation with locally adaptive dynamic models. *Comput. Statist.* **17** 479–499. MR1952693

LASSAS, M. and SILTANEN, S. (2004). Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Probl.* **20** 1537.

LEBRUN, M., BUADES, A. and MOREL, J. M. (2013). A nonlocal Bayesian image denoising algorithm. *SIAM J. Imaging Sci.* **3** 1665–1688.

LEYTON, A. S. and SHERRINGTON, C. S. (1917). Observations on the excitable cortex of the chimpanzee, orangutan, and gorilla. *Q.j. Exp. Physiol.* **11** 135–222.

LOUCHET, C. and MOISAN, L. (2013). Posterior expectation of the total variation model: Properties and experiments. *SIAM J. Imaging Sci.* **6** 2640–2684. MR3143828

MACHADO, T. A., PNEVMATIKAKIS, E., PANINSKI, L., JESSELL, T. and MIRI, A. (2015). Primacy of flexor locomotor pattern revealed by ancestral reversion of motor neuron identity. *Cell* **162** 338–350.

MACKAY, D. (1995). Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network* **6** 469–505.

MACKE, J. H., GERWINN, S., WHITE, L. E., KASCHUBE, M. and BETHGE, M. (2010). Bayesian estimation of orientation preference maps. *Adv. Neural Inf. Process. Syst.* **22** 1195–1203.

MACKE, J. H., GERWINN, S., WHITE, L. E., KASCHUBE, M. and BETHGE, M. (2011). Gaussian process methods for estimating cortical maps. *NeuroImage* **56** 570–581.

METIN, C., GODEMENT, P. and IMBERT, M. (1988). The primary visual cortex in the mouses: Receptive field properties and functional organization. *Exp. Brain Res.* **69**.

MOTWANI, M. C., GADIYA, M. C., MOTWANI, R. C. and HARRIS, F. C. (2004). Survey of image denoising techniques. In *Global Signal Processing Expo*, Santa Clara, CA.

MURPHY, E. H. and BERMAN, N. (1979). The rabit and the cat: A comparison of some features of response properties of single cells in the primary visual cortex. *J. Comp. Neurol.* **188**.

MURRAY, I., ADAMS, R. P. and MACKAY, D. (2010). Elliptical slice sampling. In *The Proceedings of the* 13*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) **9** 541–548.

OHKI, K., CHUNG, S., CH'NG, Y., KARA, P. and REID, C. (2005). Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* **433** 597–603.

OHKI, K., CHUNG, S., KARA, P., HUBENER, M., BONHOEFFER, T. and REID, R. C. (2006). Highly ordered arrangement of single neurons in orientation pinwheels. *Nature* **442** 925–928. DOI:10.1038/nature05019.

OLIVEIRA, J., BIOUCAS-DIAS, J. M. and FIGUEIREDO, M. (2009). Adaptive total variation image deblurring: A majorization-minimization approach. *Signal Process.* **89** 1683–1693.

PANINSKI, L. (2010). Fast Kalman filtering on quasilinear dendritic trees. *J. Comput. Neurosci.* **28** 211–228. MR2609429

PANINSKI, L., AHMADIAN, Y., FERREIRA, D. G., KOYAMA, S., RAHNAMA RAD, K., VIDNE, M., VOGELSTEIN, J. and WU, W. (2010). A new look at state-space models for neural data. *J. Comput. Neurosci.* **29** 107–126. MR2721336

PAPANDREOU, G. and YUILLE, A. (2010). Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems* 23 (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 1858–1866. Curran Associates, Red Hook, NY.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

PENFIELD, W. and RASMUSSEN, T. (1950). The cerebral cortex of man; a clinical study of localization of function. *J. Amer. Med. Assoc.* **144**.

PENNY, W. D., TRUJILLO-BARRETO, N. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

PNEVMATIKAKIS, E. A., GAO, Y., SOUDRY, D., PFAU, D., LACEFIELD, C., POSKANZER, K., BRUNO, R., YUSTE, R. and PANINSKI, L. (2014a). A structured matrix factorization framework for large scale calcium imaging data analysis. Preprint. Available at arXiv:1409.2903.

PNEVMATIKAKIS, E. A., RAHNAMA RAD, K., HUGGINS, J. and PANINSKI, L. (2014b). Fast Kalman filtering and forward-backward smoothing via low-rank perturbative approach. *J. Comput. Graph. Statist.* **23** 316–339. MR3215813

PORTILLA, J., STRELA, V., WAINWRIGHT, M. J. and SIMONCELLI, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12** 1338–1351. MR2026777

PORTUGUES, R., FEIERSTEIN, C. E., ENGERT, F. and ORGER, M. B. (2014). Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron* **81** 1328–1343. DOI:10.1016/j.neuron.2014.01.019.

PRESS, W., TEUKOLSKY, S., VETTERLING, W. and FLANNERY, B. (1992). *Numerical Recipes in C*. Cambridge Univ. Press, Cambridge.

PREVEDEL, R., YOON, Y., HOFFMANN, M., PAK, N., WETZSTEIN, G., KATO, S., SCHRÖDEL, T., RASKAR, R., ZIMMER, M., BOYDEN, E. et al. (2014). Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11** 727–730.

QUIROS, A., DIEZ, R. M. and GAMERMAN, D. (2010). Bayesian spatiotemporal model of fMRI data. *NeuroImage* **49** 442–456.

RAHNAMA RAD, K. and PANINSKI, L. (2010). Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods. *Network* **21** 142–168.

RASMUSSEN, C. and WILLIAMS, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

REICHL, L., LÖWEL, S. and WOLF, F. (2009). Pinwheel stabilization by ocular dominance segregation. *Phys. Rev. Lett.* **102** 208101.

RIEKE, F., WARLAND, D., DE RUYTER VAN STEVENINCK, R. and BIALEK, W. (1997). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

ROBERTS, G. O. and STRAMER, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.* **4** 337–357. MR2002247

ROMANES, G. J. (1964). The motor pools of the spinal cord. *Prog. Brain Res.* **11** 93–119.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D* **60** 259–268. MR3363401

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Taylor & Francis, London.

SCHNABEL, M., KASCHUBE, M., LOWEL, S. and WOLF, F. (2007). Random waves in the brain: Symmetries and defect generation in the visual cortex. *Eur. Phys. J. Spec. Top.* **145** 137–157.

SCOTT, S. H. (2000). Population vectors and motor cortex: Neural coding or epiphenomenon? *Nat. Neurosci.* **3** 307–308.

SHMUEL, A. and GRINVALD, A. (1996). Functional organization for direction of motion and its relationship to orientation maps in cat area 18. *J. Neurosci.* **16** 6945–6964.

SIDEN, P., EKLUND, A., BOLIN, D. and VILLANI, M. (2016). Fast Bayesian whole-brain fMRI analysis with spatial 3D priors. Preprint. Available at arXiv:1606.00980v1 [stat.CO].

SLAWSKI, M. (2012). The structured elastic net for quantile regression and support vector classification. *Stat. Comput.* **22** 153–168. MR2865062

SLAWSKI, M., ZU CASTELL, W. and TUTZ, G. (2010). Feature selection guided by structural information. *Ann. Appl. Stat.* **4** 1055–1080. MR2758433

STARCK, J., CANDES, E. and DONOHO, D. (2002). The curvevlet transform for image denoising. *IEEE Trans. Image Process.* **11** 670–684.

STEVENSON, R. and DELP, E. (1990a). Fitting curves with discontinuities. In *Proceedings of International Workshop on Robust Comput. Vision*, 127–136.

STEVENSON, R. and DELP, E. (1990b). Surface reconstruction with discontinuities. *Proc. SPIE Int. Soc. Opt. Eng.* **1610** 46–57.

SWINDALE, N. V. (1998). Orientation tuning curves: Empirical description and estimation of parameters. *Biol. Cybernet.* **78** 45–56.

SWINDALE, N. V. (2008). Visual map. *Scholarpedia* **3** 4607.

TIAO, Y. C. and BLAKEMORE, C. (1976). Functional organization in the visual cortex of the golden hamster. *J. Comp. Neurol.* **168** 459–481.

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

VAN GERVEN, M. A. J., CSEKE, B., DE LANGE, F. P. and HESKES, T. (2010). Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage* **50** 150–161.

VAN HOOSER, S. D., HEIMEL, J. A. F., CHUNG, S., NELSON, S. B. and TOTH, L. J. (2005). Orientation selectivity without orientation maps in visual cortex of a highly visual mammal. *J. Neurosci*. **25** 19–28.

VIDNE, M., AHMADIAN, Y., SHLENS, J., PILLOW, J. W., KULKARNI, J., LITKE, A. M., CHICHILNISKY, E. J., SIMONCELLI, E. and PANINSKI, L. (2012). Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *J. Comput. Neurosci*. **33** 97–121. MR2956393

VOGEL, C. R. and OMAN, M. E. (1996). Iterative methods for total variation denoising. *SIAM J. Sci. Comput*. **17** 227–238. MR1375276

VOGEL, C. R. and OMAN, M. E. (1998). Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Image Process*. **7** 813–824. MR1667392

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

WANDELL, B. (1995). *Foundations of Vision*. Sinauer, Boston, MA.

WANG, Y., YANG, J., YIN, W. and ZHANG, Y. (2009). A new alternative minimization algorithm for total variation image reconstruction. *SIAM J. Imaging Sci*. **1** 248–272.

WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res*. **17** 1–41. MR3543511

WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. MR0909372

WILSON, S. and MOORE, C. (2015). S1 somatotopic maps. *Scholarpedia* **10** 8574.

WIPF, D. and NAGARAJAN, S. (2008). A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems* 1625–1632. Curran Associates, Red Hook.

WOOLRICH, M. W. (2012). Bayesian inference in fMRI. *NeuroImage* **62** 801–810.

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag*. **23** 213–231.

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

YUE, Y., LOH, J. M. and LINDQUIST, M. A. (2010). Adaptive spatial smoothing of fMRI images. *Stat. Interface* **3** 3–13. MR2609707

YUE, Y. and SPECKMAN, P. L. (2010). Nonstationary spatial Gaussian Markov random fields. *J. Comput. Graph. Statist*. **19** 96–116. MR2654402

YUE, Y. R., SPECKMAN, P. L. and SUN, D. (2012). Priors for Bayesian adaptive spline smoothing. *Ann. Inst. Statist. Math*. **64** 577–613. MR2880870

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 638–654
DOI: 10.1214/16-AOAS997

IMPROVING EFFICIENCY IN BIOMARKER INCREMENTAL VALUE EVALUATION UNDER TWO-PHASE DESIGNS[1]

BY YINGYE ZHENG[*], MARSHALL BROWN[*], ANNA LOK[†] AND TIANXI CAI[‡]

Fred Hutchinson Cancer Research Center[*], University of Michigan[†] and Harvard T.H. Chan School of Public Health[‡]

Cost-effective yet efficient designs are critical to the success of biomarker evaluation research. Two-phase sampling designs, under which expensive markers are only measured on a subsample of cases and noncases within a prospective cohort, are useful in novel biomarker studies for preserving study samples and minimizing cost of biomarker assaying. Statistical methods for quantifying the predictiveness of biomarkers under two-phase studies have been proposed [*Biostatistics* **13** (2012) 89–100, *Biometrics* **68** (2012) 1219–1227]. These methods are based on a class of inverse probability weighted (IPW) estimators where weights are "true" sampling weights that simply reflect the sampling strategy of the study. While simple to implement, existing IPW estimators are limited by lack of practicality and efficiency. In this manuscript, we investigate a variety of two-phase design options and provide statistical approaches aimed at improving the efficiency of simple IPW estimators by incorporating auxiliary information available for the entire cohort. We consider accuracy summary estimators that accommodate auxiliary information in the context of evaluating the incremental values of novel biomarkers over existing prediction tools. In addition, we evaluate the relative efficiency of a variety of sampling and estimation options under two-phase studies, shedding light on issues pertaining to both the design and analysis of biomarker validation studies. We apply our methods to the evaluation of a novel biomarker for liver cancer risk conducted with a two-phase nested case control design [*Gastroenterology* **138** (2010) 493–502].

REFERENCES

BORGAN, Ø., GOLDSTEIN, L. and LANGHOLZ, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23** 1749–1778. MR1370306

BORGAN, Ø., LANGHOLZ, B., SAMUELSEN, S. O., GOLDSTEIN, L. and POGODA, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6** 39–58. MR1767493

BRESLOW, N. E., DAY, N. E. et al. (1980). *Statistical Methods in Cancer Research*, *Vol.* 1. IARC Publications.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009a). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009b). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences* **1** 32–49.

Key words and phrases. Biomarker, prediction accuracy, risk prediction, two-phase study.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201. MR1892660

CAI, T. and ZHENG, Y. (2011). Nonparametric evaluation of biomarker accuracy under nested case-control studies. *J. Amer. Statist. Assoc.* **106** 569–580. MR2847971

CAI, T. and ZHENG, Y. (2012). Evaluating prognostic accuracy of biomarkers under nested case-control studies. *Biostatistics* **13** 89–100.

CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845.

CHENG, S. C., WEI, L. J. and YING, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *J. Amer. Statist. Assoc.* **92** 227–235. MR1436111

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc., B* **34** 187–220. MR0341758

GRAY, R. J. (2009). Weighted analyses for cohort sampling designs. *Lifetime Data Anal.* **15** 24–40.

LANGHOLZ, B. and BORGAN, Y. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53** 767–774.

LIU, D., CAI, T. and ZHENG, Y. (2012). Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics* **68** 1219–1227.

LOK, A. S., STERLING, R. K., EVERHART, J. E., WRIGHT, E. C., HOEFS, J. C., DI BISCEGLIE, A. M., MORGAN, T. R., KIM, H.-Y., LEE, W. M., BONKOVSKY, H. L. et al. (2010). Des-γ-carboxy prothrombin and α-fetoprotein as biomarkers for the early detection of hepatocellular carcinoma. *Gastroenterology* **138** 493–502.

MURPHY, S. A., ROSSINI, A. J. and VAN DER VAART, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127

PENCINA, M. J., D'AGOSTINO, R. B. SR., D'AGOSTINO, R. B. JR. and VASAN, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **27** 157–172. MR2412695

PFEIFFER, R. M. and GAIL, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67** 1057–1065. MR2829240

PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.

QI, L., WANG, C. Y. and PRENTICE, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* **100** 1250–1263. MR2236439

SAEGUSA, T. and WELLNER, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.* **41** 269–295. MR3059418

SAMUELSEN, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84** 379–394.

SELF, S. G., PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16** 64–81. MR0924857

THOMAS, D. C. (1977). Addendum to "Methods of cohort analysis: Appraisal by application to asbestos mining." *J. Roy. Statist. Soc. Ser. A* **140** 483–485.

WANG, S. and WANG, C. Y. (2001). A note on kernel assisted estimators in missing covariate regression. *Statist. Probab. Lett.* **55** 439–449. MR1877649

WANG, T., ROHAN, T. E., GUNTER, M. J., XUE, X., WACTAWSKI-WENDE, J., RAJPATHAK, S. N., CUSHMAN, M., STRICKLER, H. D., KAPLAN, R. C., WASSERTHEIL-SMOLLER, S., SCHERER, P. E. and GLORIA, Y. F. HO (2011). A prospective study of inflammation markers and endometrial cancer risk in postmenopausal hormone nonusers. *Cancer Epidemiol. Biomark. Prev.* **20** 971–977.

ZENG, D. and LIN, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93** 627–640. MR2261447

ZENG, D. and LIN, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Amer. Statist. Assoc.* **109** 371–383. MR3180570

ZHENG, Y., BROWN, M., LOK, A. and CAI, T. (2017). Supplement to "Improving Efficiency in Biomarker Incremental Value Evaluation under Two-phase Designs." DOI:10.1214/16-AOAS997SUPP.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 655–679
DOI: 10.1214/16-AOAS1011

STOCHASTIC MODELLING AND INFERENCE IN ELECTRONIC HOSPITAL DATABASES FOR THE SPREAD OF INFECTIONS: *CLOSTRIDIUM DIFFICILE* TRANSMISSION IN OXFORDSHIRE HOSPITALS 2007–2010[1]

BY MADELEINE CULE AND PETER DONNELLY[2]

University of Oxford

The combination of genetic information with electronic patient records promises to provide a powerful new resource for understanding human disease and its treatment. Here we develop and apply a novel stochastic compartmental model to a large dataset on *Clostridium difficile* infection (CDI) in three Oxfordshire hospitals over a 2.5 year period which combines genetic information on 858 confirmed cases of CDI with a database of 750,000 patient records. *C. difficile* is a major cause of healthcare-associated diarrhoea and is responsible for substantial mortality and morbidity, with relatively little known about its biology or its transmission epidemiology. Bayesian analysis of our model, via Markov chain Monte Carlo, provides new information about the biology of CDI, including genetic heterogeneity in infectiousness across different sequence types, and evidence for ward contamination as a significant mode of transmission, and allows inferences about the contribution of particular individuals, wards or hospitals to transmission of the bacterium, and assessment of changes in these over time following changes in hospital practice. Our work demonstrates the value of using statistical modelling and computational inference on large-scale hospital patient databases and genetic data.

REFERENCES

BARTLETT, J. G., CHANG, T. E. W., GURWITH, M., GORBACH, S. L. and ONDERDONK, A. B. (1978). Antibiotic-associated pseudomembranous colitis due to toxin-producing clostridia. *N. Engl. J. Med.* **298** 531–534.

BEST, E. L., FAWLEY, W. N., PARNELL, P. and WILCOX, M. H. (2010). The potential for airborne dispersal of *Clostridium difficile* from symptomatic patients. *Clin. Infect. Dis.* **50** 1450–1457.

BOBULSKY, G. S., AL-NASSIR, W. N., RIGGS, M. M., SETHI, A. K. and DONSKEY, C. J. (2008). *Clostridium difficile* skin contamination in patients with *C. difficile*-associated disease. *Clin. Infect. Dis.* **46** 447–450.

BOURDAIN, A. (2001). *Typhoid Mary*: *An Urban Historical*. Bloomsbury, New York.

COOPER, B. S., MEDLEY, G. F., BRADLEY, S. J. and SCOTT, G. M. (2008). An augmented data method for the analysis of nosocomial infection data. *Am. J. Epidemiol.* **168** 548–557.

COOPER, B. S., KYPRAIOS, T., BATRA, R., WYNCOLL, D., TOSAS, O. and EDGEWORTH, J. D. (2012). Quantifying type-specific reproduction numbers for nosocomial pathogens: Evidence for heightened transmission of an Asian sequence type 239 MRSA clone. *PLoS Comput. Biol.* **8** e1002454.

Key words and phrases. Stochastic modelling, Markov chain Monte Carlo, medicine.

CULE, M. and DONNELLY, P. (2017). Supplement to "Stochastic modelling and inference in electronic hospital databases for the spread of infections: *Clostridium difficile* transmission in Oxfordshire hospitals 2007–2010." DOI:10.1214/16-AOAS1011SUPP.

DONSKEY, C. J. (2010). Preventing transmission of *Clostridium difficile*: Is the answer blowing in the wind? *Clin. Infect. Dis.* **50** 1458–1461.

FINNEY, J. M., WALKER, A. S., PETO, T. E. A. and WYLLIE, D. H. (2011). An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med. Inform. Decis. Mak.* **11** 7.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

GRIFFITHS, D., FAWLEY, W., KACHRIMANIDOU, M., BOWDEN, R., CROOK, D. W., FUNG, R., GOLUBCHIK, T., HARDING, R. M., JEFFERY, K. J. M., JOLLEY, K. A., KIRTON, R., PETO, T. E. A., REES, G., STOESSER, N., VAUGHAN, A., WALKER, A. S., YOUNG, B. C., WILCOX, M. and DINGLE, K. E. (2010). Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48** 770–778.

HEALTH PROTECTION AGENCY (2009). Voluntary surveillance of *Clostridium difficile* in England, Wales and Northern Ireland, 2008.

HEALTH PROTECTION AGENCY (2010). *Clostridium difficile* Ribotyping Network (CDRN) for England and Northern Ireland Annual Report 2010/2011.

JOHNSON, S. (2009). Recurrent *Clostridium difficile* infection: A review of risk factors, treatments, and outcomes. *J. Infect.* **58** 403–410.

KARAS, J. A., ENOCH, D. A. and ALIYU, S. H. (2010). A review of mortality due to *Clostridium difficile* infection. *J. Infect.* **61** 1–8.

KYPRAIOS, T., NEILL, P. D. O., HUANG, S. S., RIFAS-SHIMAN, S. L. and COOPER, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infect. Dis.*

LANZAS, C., DUBBERKE, E. R., LU, Z., RESKE, K. A. and GRÖHN, Y. T. (2011). Epidemiological model for *Clostridium difficile* transmission in healthcare settings. *Infect. Control Hosp. Epidemiol.* **32** 553–561.

LLOYD-SMITH, J. O., SCHREIBER, S. J., KOPP, P. E. and GETZ, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature* **438** 355–359.

LOO, V. G., BOURGAULT, A.-M., POIRIER, L., LAMOTHE, F., MICHAUD, S., TURGEON, N., TOYE, B., BEAUDOIN, A., FROST, E. H., GILCA, R., BRASSARD, P., DENDUKURI, N., BÉLIVEAU, C., OUGHTON, M., BRUKNER, I. and DASCAL, A. (2011). Host and pathogen factors for *Clostridium difficile* infection and colonization. *N. Engl. J. Med.* **365** 1693–1703.

MCDONALD, L. L. C., KILLGORE, G. E., THOMPSON, A., OWENS JR, R. C., KAZAKOVA, S. V., SAMBOL, S. P., JOHNSON, S. and GERDING, D. N. (2005). An epidemic, toxin gene-variant strain of *Clostridium difficile*. *N. Engl. J. Med.* **353** 2433–2441.

MCFARLAND, L. V., MULLIGAN, M. E., KWOK, R. Y. Y. and STAMM, W. E. (1989). Nosocomial acquisition of *Clostridium difficile* infection. *N. Engl. J. Med.* **320** 204–210.

PLANCHE, T., AGHAIZU, A., HOLLIMAN, R., RILEY, P., POLONIECKI, J., BREATHNACH, A. and KRISHNA, S. (2008). Diagnosis of *Clostridium difficile* infection by toxin detection kits: A systematic review. *Lancet, Infect. Dis.* **8** 777–784.

ROLFE, R. D. (1988). Asymptomatic intestinal colonization by *Clostridium difficile*. In *Clostridium Difficile*: *Its Role in Intestinal Disease* 201–225. Academic Press, San Diego, CA.

STARR, J. M. and CAMPBELL, A. (2001). Mathematical modeling of *Clostridium difficile* infection. *Clin. Microbiol. Infect.* **7** 432–437.

STARR, J. M., CAMPBELL, A., RENSHAW, E., POXTON, I. R. and GIBSON, G. J. (2009). Spatio-temporal stochastic modelling of *Clostridium difficile*. *J. Hosp. Infect.* **71** 49–56.

WALKER, A. S., EYRE, D. W., WYLLIE, D. H., DINGLE, K. E., HARDING, R. M., O'CONNOR, L., GRIFFITHS, D., VAUGHAN, A., FINNEY, J. M., WILCOX, M. H., CROOK, D. W., PETO, T. E. A. and WALKER, A. S. (2011). Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Med*. **9** e1001172.

WILCOX, M. and FAWLEY, W. (2000). Hospital disinfectants and spore formation by *Clostridium difficile*. *Lancet* **356** 1324–1324.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 680–697
DOI: 10.1214/16-AOAS1012

MODELING LOG-LINEAR CONDITIONAL PROBABILITIES FOR ESTIMATION IN SURVEYS

BY YVES THIBAUDEAU*, ERIC SLUD*,† AND ALFRED GOTTSCHALCK*

*U.S. Census Bureau** *and University of Maryland*†

The Survey of Income and Program Participation (SIPP) is a survey with a longitudinal structure and complex nonignorable design, for which correct estimation requires using the weights. The longitudinal setting also suggests conditional-independence relations between survey variables and early- versus late-wave employment classifications. We state original assumptions justifying an extension of the partially model-based approach of Pfeffermann, Skinner and Humphreys [*J. Roy. Statist. Soc. Ser. A* **161** (1998) 13–32], accounting for the design of SIPP and similar longitudinal surveys. Our assumptions support the use of log-linear models of longitudinal survey data. We highlight the potential they offer for simultaneous bias-control and reduction of sampling error relative to direct methods when applied to small subdomains and cells. Our assumptions allow us to innovate by showing how to rigorously use only a longitudinal survey to estimate a complex log-linear longitudinal association structure and embed it in cross-sectional totals to construct estimators that can be more efficient than direct estimators for small cells.

REFERENCES

ABOWD, J., STEPHENS, B., VILHUBER, L., ANDERSSON, F., MCKINNEY, K., ROEMER, M. and WOODCOCK, S. (2005). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. Technical Paper TP 2006-01. Available at lehd.ces.census.gov/doc/technical_paper/tp-2006-01.pdf.

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR3087436

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. MR0381130

CHAMBERS, R. L. FEENEY, G. A. (1977). Log linear models for small area estimation. Paper presented at the Joint Conference of the CSIRO Division of Mathematics and the Australian of the Biometrics Society, Newcastle, Australia, Biometrics Abstract 2655.

CONAWAY, M. and LOHR, L. (1994). A longitudinal analysis of factors associated with reporting violent crimes to the police. *J. Quant. Criminol.* **10** 23–39.

FIENBERG, S. E. (1980). The measurement of crime victimization: Prospect for panel analysis of a panel survey. *J. Roy. Statist. Soc. Ser. D* **29** 313–350.

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

Key words and phrases. Log-linear model, conditional probability, Horvitz–Thompson estimator, model calibration.

FIENBERG, S. E. and STASNY, E. (1983). Estimating monthly gross flows in labour force participation. *Surv. Methodol.* **9** 78–101.

FULLER, W. (2009). *Sampling Statistics*. Wiley, New York.

FULLER, W. and ISAKI, C. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling* (D. Krewski, J. N. K. Rao and R. Platek, eds.). Academic Press, San Diego.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

GHOSH, M. and RAO, J. N. K. (1994). Small area estimation: An appraisal. *Statist. Sci.* **9** 55–93. MR1278679

HABER, S. (1985). Applications of a Matched File Linking the Bureau of the Census Survey of Income Program Participation and Income Data. The Survey of Income and Program Participation WP 3.

JUDKINS, D. (1990). Fay's method for variance estimation. *J. Off. Stat.* **6** 223–239.

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9** 1010–1019. MR0628756

LOPEZ-VIZCAINO, E., LOMBARDÍA, M. and MORALES, D. (2015). Small area estimation of labour force indicators under a multinomial model with correlated time and area effects. *J. Roy. Statist. Soc. Ser. A* **178** 535–575.

MARKER, D. (1999). Organization of small area estimates estimators using a generalized linear regression framework. *J. Off. Stat.* **15** 1–24.

MOLINA, I., SAEI, A. and LOMBARDÍA, M. J. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *J. Roy. Statist. Soc. Ser. A* **170** 975–1000. MR2408988

NOBLE, A., HASLETT, S. and ARNOLD, G. (2002). Small area estimation via generalized linear models. *J. Off. Stat.* **18** 45–60.

PFEFFERMANN, D., SKINNER, C. and HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *J. Roy. Statist. Soc. Ser. A* **161** 13–32.

PURCELL, N. J. and KISH, L. (1979). Estimation for small domains. *Biometrics* **35** 365–384. MR0535774

PURCELL, N. J. and KISH, L. (1980). Postcensal estimates from local areas (or domains). *Int. Stat. Rev.* **48** 3–18.

PURCELL, N. J. (1979). Efficient Small Domain Estimation: A Categorical Data Approach. Unpublished Ph.D. Thesis, Univ. Michigan.

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

RAO, J. N. K. and SCOTT, A. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Ann. Statist.* **12** 46–60. MR0733498

REILLY, C., GELMAN, A. and KATZ, J. (2001). Poststratification without population level information on the poststratifying variable with application to political polling. *J. Amer. Statist. Assoc.* **96** 1–11. MR1973779

SAPHIRE, D. (1984). *Estimation of Victimization Prevalence Using Data from the National Crime Survey*. Springer, New York.

SCOTT, J. (2005). SIPP2004+: Cross-Sectional Weighting Specifications for Wave 1. Memorandum WGT-20. U.S. Department of Commerce, U.S. Census Bureau, Washington, DC.

SKINNER, C. (2011). Log-linear Modelling with Complex Survey Data. Proceedings 58th World Statistical Congress, 2011, Dublin (Session IPS056).

SLAVKOVIĆ, A., ZHU, X. and PETROVIĆ, S. (2015). Fibers of multi-way contingency tables given conditionals: Relation to marginals, cell bounds and Markov bases. *Ann. Inst. Statist. Math.* **67** 621–648. MR3357932

STASNY, E. (1987). Some Markov-chain models for nonresponse in estimating Gross labor force flows. *J. Off. Stat.* **3** 359–373.

THIBAUDEAU, Y., SLUD, E. and GOTTSCHALCK, A. (2017). Supplement to "Modeling log-linear conditional probabilities for estimation in surveys." DOI:10.1214/16-AOAS1012SUPP.

U.S. BUREAU OF THE CENSUS (2009). SIPP User Guide's Sample Design and Interview Procedures Chapter 2. Available at http://www2.census.gov/programs-surveys/sipp/guidance/SIPP_2008_USERS_Guide_Chapter2.pdf.

WINKLER, W. (1993). On Dykstra's iterative fitting procedure. *Ann. Probab.* **18** 1410–1415. MR1062075

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. Springer, New York. MR0799715

YBARRA, L. and LOHR, S. (2002). Estimates of repeat victimization using the national crime victimization survey. *J. Quant. Criminol.* **18** 1–21.

ZHANG, L.-C. and CHAMBERS, R. L. (2004). Small area estimates for cross-classifications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 479–496. MR2062389

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 698–724
DOI: 10.1214/16-AOAS1013

PHOTO-z ESTIMATION: AN EXAMPLE OF NONPARAMETRIC CONDITIONAL DENSITY ESTIMATION UNDER SELECTION BIAS[1]

BY RAFAEL IZBICKI*, ANN B. LEE† AND PETER E. FREEMAN†

Federal University of São Carlos and Carnegie Mellon University†*

Redshift is a key quantity for inferring cosmological model parameters. In photometric redshift estimation, cosmologists use the coarse data collected from the vast majority of galaxies to predict the redshift of individual galaxies. To properly quantify the uncertainty in the predictions, however, one needs to go beyond standard regression and instead estimate the *full conditional density* $f(z|\mathbf{x})$ of a galaxy's redshift z given its photometric covariates $\mathbf{x}$. The problem is further complicated by *selection bias*: usually only the rarest and brightest galaxies have known redshifts, and these galaxies have characteristics and measured covariates that do not necessarily match those of more numerous and dimmer galaxies of unknown redshift. Unfortunately, there is not much research on how to best estimate complex multivariate densities in such settings.

Here we describe a general framework for properly constructing and assessing nonparametric conditional density estimators under selection bias, and for combining two or more estimators for optimal performance. We propose new improved photo-z estimators and illustrate our methods on data from the Sloan Data Sky Survey and an application to galaxy–galaxy lensing. Although our main application is photo-z estimation, our methods are relevant to any high-dimensional regression setting with complicated asymmetric and multimodal distributions in the response variable.

REFERENCES

AIHARA, H. et al. (2011). The eighth data release of the Sloan Digital Sky Survey: First data from SDSS-III. *Astrophys. J., Suppl. Ser.* **193** 29.

BALL, N. M. and BRUNNER, R. J. (2010). Data mining and machine learning in astronomy. *Internat. J. Modern Phys. D* **19** 1049–1106.

BICKEL, S., BRÜCKNER, M. and SCHEFFER, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.* **10** 2137–2155. MR2550104

CORRADI, V. and SWANSON, N. R. (2006). Predictive density evaluation. In *Handbook of Economic Forecasting*. North-Holland, Amsterdam.

CUNHA, C. E., LIMA, M., OYAIZU, H., FRIEMAN, J. and LIN, H. (2009). Estimating the redshift distribution of photometric galaxy samples—II. Applications and tests of a new method. *Mon. Not. R. Astron. Soc.* **396** 2379–2398.

DAHLEN, T., MOBASHER, B., FABER, S. M., FERGUSON, H. C., BARRO, G., FINKELSTEIN, S. L., FINLATOR, K., FONTANA, A., GRUETZBAUCH, R., JOHNSON, S. et al. (2013).

Key words and phrases. Density estimation, nonparametric statistics, selection bias, photometric redshift.

A critical assessment of photometric redshift methods: A CANDELS investigation. *Astrophys. J.* **775** 93.

FERNÁNDEZ-SOTO, A., LANZETTA, K. M. and YAHIL, A. (1998). A new catalog of photometric redshifts in the Hubble Deep Field. *Astrophys. J.* **513** 34–50.

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLKOPF, B. (2010). Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning* (J. Quionero-Candela, M. Sugiyama, A. Schwaighofer and N. D. Lawrence, eds.) Chapter 8. MIT Press, Cambridge, MA.

HALL, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.* **15** 1491–1519. MR0913570

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Berlin.

IZBICKI, R. and LEE, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *J. Comput. Graph. Statist.* **25** 1297–1316. MR3572041

IZBICKI, R., LEE, A. B and FREEMAN, P. E (2017). Supplement to "Photo-z estimation: An example of nonparametric conditional density estimation under selection bias." DOI:10.1214/16-AOAS1013SUPP.

IZBICKI, R., LEE, A. B. and SCHAFER, C. M. (2014). High-dimensional density ratio estimation with extensions to approximate likelihood computation. *J. Mach. Learn. Res.* **33**.

KANAMORI, T., HIDO, S. and SUGIYAMA, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* **10** 1391–1445. MR2534866

KANAMORI, T., SUZUKI, T. and SUGIYAMA, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Mach. Learn.* **86** 335–367. MR2897527

KIND, M. C. and BRUNNER, R. J. (2013). Tpz: Photometric redshift pdfs and ancillary information by using prediction trees and random forests. *Mon. Not. R. Astron. Soc.* **432** 1483–1501.

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KREMER, J., GIESEKE, F., PEDERSEN, K. S. and IGEL, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astron. Comput.*

LEE, A. B. and IZBICKI, R. (2016). A spectral series approach to high-dimensional nonparametric regression. *Electron. J. Stat.* **10** 423–463. MR3466189

LIMA, M., CUNHA, C. E., OYAIZU, H., FRIEMAN, J., LIN, H. and SHELDON, E. (2008). Estimating the redshift distribution of photometric galaxy samples. *Mon. Not. R. Astron. Soc.* **390** 118–130.

LOOG, M. (2012). Nearest neighbor-based importance weighting. In *IEEE International Workshop on Machine Learning for Signal Processing*.

MANDELBAUM, R., SELJAK, U., HIRATA, C. M., BARDELLI, S., BOLZONELLA, M., BONGIORNO, A., CAROLLO, M., CONTINI, T., CUNHA, C. E., GARILLI, B., IOVINO, A., KAMPCZYK, P., KNEIB, J. P., KNOBEL, C., KOO, D. C., LAMAREILLE, F., LE FEVRE, O., LEBORGNE, J. F., LILLY, S. J., MAIER, C., MAINIERI, V., MIGNOLI, M., NEWMAN, J. A., OESCH, P. A., PEREZ-MONTERO, E., RICCIARDELLI, E., SCODEGGIO, M., SILVERMAN, J. and TASCA, L. (2008). Precision photometric redshift calibration for galaxy–galaxy weak lensing. *Mon. Not. R. Astron. Soc.* **386** 781–806.

MARGOLIS, A. (2011). A literature review of domain adaptation with unlabeled data. Available at http://ssli.ee.washington.edu/~amargoli/review_Mar23.pdf.

MINH, H. Q., NIYOGI, P. and YAO, Y. (2006). Mercer's theorem, feature maps, and smoothing. In *Learning Theory*, 19*th Annual Conference on Learning Theory*.

MORENO-TORRES, J. G., RAEDER, T., ALAÍZ-RODRÍGUEZ, R., CHAWLA, N. V. and HERRERA, F. (2012). A unifying view on dataset shift in classification. *Pattern Recogn.* **45** 521–530.

OYAIZU, H., LIMA, M., CUNHA, C. E., LIN, H. and FRIEMAN, J. (2008). Photometric redshift error estimators. *Astrophys. J.* **689** 709–720.

QUIONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

SHELDON, E. S., CUNHA, C. E., MANDELBAUM, R., BRINKMANN, J. and WEAVER, B. A. (2012). Photometric redshift probability distributions for galaxies in the SDSS DR8. *Astrophys. J., Suppl. Ser.* **201** 32.

SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598

SPRINGEL, V., FRENK, C. S. and WHITE, S. D. M. (2006). The large-scale structure of the universe. *Nature* **440** 1137–1144.

SUGIYAMA, M., SUZUKI, T., NAKAJIMA, S., KASHIMA, H., VON BÜNAU, P. and KAWANABE, M. (2008). Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.* **60** 699–746. MR2453568

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York.

WEINER, B. J., PHILLIPS, A. C., FABER, S. M., WILLMER, C. N. A., VOGT, N. P. et al. (2005). The DEEP groth strip galaxy redshift survey. III. Redshift catalog and properties of galaxies. *Astrophys. J.* **620** 595.

WITTMAN, D. (2009). What lies beneath: Using $p(z)$ to reduce systematic photometric redshift errors. *Astrophys. J. Lett.* **700**.

YORK, D. G., ADELMAN, J., ANDERSON, J. E. JR., ANDERSON, S. F., ANNIS, J., BAHCALL, N. A., BAKKEN, J. A., BARKHOUSER, R., BASTIAN, S., BERMAN, E. et al. (2000). The Sloan Digital Sky Survey: Technical summary. *Astrophys. J.* **120** 1579.

ZHAO, L. C. and LIU, Z. J. (1985). Strong consistency of the kernel estimators of conditional density function. *Acta Math. Appl. Sin. Engl. Ser.* **1** 314–318. MR0867902

ZHENG, H. and ZHANG, Y. (2012). Review of techniques for photometric redshift estimation. In *Software and Cyberinfrastructure for Astronomy II* **8451**.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 725–750
DOI: 10.1214/16-AOAS1015

HYPOTHESIS TESTING FOR NETWORK DATA IN FUNCTIONAL NEUROIMAGING

BY CEDRIC E. GINESTET[*,1], JUN LI[†,2], PRAKASH BALACHANDRAN[†], STEVEN ROSENBERG[†,2] AND ERIC D. KOLACZYK[†,1,2]

King's College London[*] and Boston University[†]

In recent years, it has become common practice in neuroscience to use networks to summarize relational information in a set of measurements, typically assumed to be reflective of either functional or structural relationships between regions of interest in the brain. One of the most basic tasks of interest in the analysis of such data is the testing of hypotheses, in answer to questions such as "Is there a difference between the networks of these two groups of subjects?" In the classical setting, where the unit of interest is a scalar or a vector, such questions are answered through the use of familiar two-sample testing strategies. Networks, however, are not Euclidean objects, and hence classical methods do not directly apply. We address this challenge by drawing on concepts and techniques from geometry and high-dimensional statistical inference. Our work is based on a precise geometric characterization of the space of graph Laplacian matrices and a nonparametric notion of averaging due to Fréchet. We motivate and illustrate our resulting methodologies for testing in the context of networks derived from functional neuroimaging data on human subjects from the 1000 Functional Connectomes Project. In particular, we show that this global test is more statistically powerful than a mass-univariate approach. In addition, we have also provided a method for visualizing the individual contribution of each edge to the overall test statistic.

REFERENCES

ACHARD, S., SALVADOR, R., WHITCHER, B., SUCKLING, J. and BULLMORE, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* **26** 63–72.

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028

AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. S. (2009). A principal component analysis for trees. *Ann. Appl. Stat.* **3** 1597–1615. MR2752149

BARDEN, D., LE, H. and OWEN, M. (2013). Central limit theorems for Frechet means in the space of phylogenetic trees. *Electron. J. Probab.* **18** 1–25. MR3035753

BECKMANN, C. F., DELUCA, M., DEVLIN, J. T. and SMITH, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **360** 1001–1013.

Key words and phrases. Fréchet mean, fMRI, graph Laplacian, hypothesis testing, matrix manifold, network data, object data.

BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

BHATIA, R. (2007). *Positive Definite Matrices*. Princeton Univ. Press, Princeton, NJ. MR2284176

BHATTACHARYA, A. and BHATTACHARYA, R. (2012). *Nonparametric Inference on Manifolds with Applications to Shape Spaces*. Cambridge Univ. Press, New York.

BHATTACHARYA, R. and LIN, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Amer. Math. Soc.* **145** 413–428. MR3565392

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498

BHATTACHARYA, R. and PATRANGENARU, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.* **33** 1225–1259. MR2195634

BHATTACHARYA, R., BUIBAS, M., DRYDEN, I., ELLINGSON, L., GROISSER, D., HENDRIKS, H., HUCKEMANN, S., LE, H., LIU, X. and MARRON, J. (2011). Extrinsic data analysis on sample spaces with a manifold stratification. In *Advances in Mathematics, Invited Contributions at the Seventh Congress of Romanian Mathematicians, Brasov* 148–156.

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931

BISWAL, B. B., MENNES, M., ZUO, X.-N., GOHEL, S. and KELLY, C. et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. USA* **107** 4734–4739.

BONNABEL, S. and SEPULCHRE, R. (2009). Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM J. Matrix Anal. Appl.* **31** 1055–1070.

BOOKSTEIN, F. (1978). *The Measurement of Biological Shape and Shape Change*. Springer, London.

BUCKNER, R. L., ANDREWS-HANNA, J. R. and SCHACTER, D. L. (2008). The brain's default network: Anatomy, function and relevance to disease. *Ann. N.Y. Acad. Sci.* **1124** 1–38.

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198.

BULLMORE, E. and SPORNS, O. (2012). The economy of brain network organization. *Nat. Rev., Neurosci.* **13** 336–349.

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684.

CAI, T., LIU, W. and LUO, X. (2011). A constrained L_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CHAVEL, I. (1984). *Eigenvalues in Riemannian Geometry*. *Pure and Applied Mathematics* **115**. Academic Press, Inc., Orlando, FL. Including a chapter by Burton Randol. With an appendix by Jozef Dodziuk. MR0768584

CHENG, S. H. and HIGHAM, N. J. (1998). A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Anal. Appl.* **19** 1097–1110. MR1636528

CHUNG, F. R. K. (1997). *Spectral Graph Theory*. *CBMS Regional Conference Series in Mathematics* **92**. Amer. Math. Soc., Providence, RI. MR1421568

DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123.

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

ELLEGREN, H. and PARSCH, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev., Genet.* **8** 689–698.

FISHER, R. (1953). Dispersion on a sphere. *Proc. R. Soc. Lond. Ser. A* **217** 295–305. MR0056866
FISHER, N. I., LEWIS, T. and EMBLETON, B. J. J. (1987). *Statistical Analysis of Spherical Data*. Cambridge Univ. Press, Cambridge. MR0899958
FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464
FU, Y. and MA, Y. (2013). *Graph Embedding for Pattern Analysis*. Springer, New York.
GINESTET, C. E., FOURNEL, A. P. and SIMMONS, A. (2014). Statistical network analysis for functional MRI: Summary networks and group comparisons. *Front. Comput. Neurosci.* **8** Art. ID 51.
GINESTET, C. E. and SIMMONS, A. (2011). Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *NeuroImage* **5** 688–704.
GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, P. and KOLACZYK, E. D. (2017). Supplement to "Hypothesis testing for network data in functional neuroimaging." DOI:10.1214/16-AOAS1015SUPP.
GREICIUS, M. D., KRASNOW, B., REISS, A. L. and MENON, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* **100** 253–258.
GROMOV, M. (2007). *Metric Structures for Riemannian and Non-Riemannian Spaces*, English ed. Birkhäuser, Inc., Boston, MA. MR2307192
HIGHAM, N. J. (2002). Computing the nearest correlation matrix: A problem from finance. *IMA J. Numer. Anal.* **22** 329–343.
HOTZ, T., HUCKEMANN, S., LE, H., MARRON, J. S., MATTINGLY, J. C., MILLER, E., NOLEN, J., OWEN, M., PATRANGENARU, V. and SKWERER, S. (2013). Sticky central limit theorems on open books. *Ann. Appl. Probab.* **23** 2238–2258. MR3127934
KANG, H., OMBAO, H., LINKLETTER, C., LONG, N. and BADRE, D. (2012). Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. *J. Amer. Statist. Assoc.* **107** 568–577. MR2980068
KENDALL, D. G. (1977). The diffusion of shape. *Adv. in Appl. Probab.* **9** 428–430.
KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. MR0737237
KENDALL, W. S. and LE, H. (2011). Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Braz. J. Probab. Stat.* **25** 323–352.
KRISHNAMACHARI, R. T. and VARANASI, M. K. (2013). On the geometry and quantization of manifolds of positive semi-definite matrices. *IEEE Trans. Signal Process.* **61** 4587–4599. MR3096701
LE, H. (2001). Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.* **33** 324–338. MR1842295
LE, H. and KUME, A. (2000). The Fréchet mean shape and the shape of the means. *Adv. in Appl. Probab.* **32** 101–113.
LEE, J. (2006). *Introduction to Smooth Manifolds*. Springer, London.
LEON, P. S., KNOCK, S. A., WOODMAN, M. M., DOMIDE, L., MERSMANN, J., MCINTOSH, A. R. and JIRSA, V. (2013). The Virtual Brain: A simulator of primate brain network dynamics. *Front. Neuroinform.* **7** Art. ID 10.
LINIAL, N. (2002). Finite metric spaces: Combinatorics, geometry and algorithms. In *Proceedings of the Eighteenth Annual Symposium on Computational Geometry* 63.
LINIAL, N., LONDON, E. and RABINOVICH, Y. (1995). The geometry of graphs and some of its algorithmic applications. *Combinatorica* **15** 215–245.
MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. Wiley, Cichester.
MCEWEN, B. S. (1999). Permanence of brain sex differences and structural plasticity of the adult brain. *Proc. Natl. Acad. Sci. USA* **96** 7128–7130.

MICHELOYANNIS, S., VOURKAS, M., TSIRKA, V., KARAKONSTANTAKI, E., KANATSOULI, K. and STAM, C. J. (2009). The influence of ageing on complex brain networks: A graph theoretical analysis. *Hum. Brain Mapp.* **30** 200–208.

MOAKHER, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **26** 735–747. MR2137480

MOAKHER, M. and ZÉRAÏ, M. (2011). The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *J. Math. Imaging Vision* **40** 171–187. MR2782125

NEWMAN, M. E. J. (2010). *Networks*: *An Introduction*. Oxford Univ. Press, Oxford. MR2676073

PACHOU, E., VOURKAS, M., SIMOS, P., SMIT, D., STAM, C., TSIRKA, V. and MICHELOYANNIS, S. (2008). Working memory in schizophrenia: An EEG study using power spectrum and coherence analysis to estimate cortical activation and network behavior. *Brain Topogr.* **21** 128–137.

SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. ID 32.

SKWERER, S., BULLITT, E., HUCKEMANN, S., MILLER, E., OGUZ, I., OWEN, M., PATRANGENARU, V., PROVAN, S. and MARRON, J. (2014). Tree-oriented analysis of brain artery structure. *J. Math. Imaging Vision* **50** 126–143.

THIRION, B., FLANDIN, G., PINEL, P., ROCHE, A., CIUCIU, P. and POLINE, J.-B. (2006). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* **27** 678–693.

TOMASI, D. and VOLKOW, N. D. (2010). Functional connectivity density mapping. *Proc. Natl. Acad. Sci. USA* **107** 9885–9890.

TOMASI, D. and VOLKOW, N. D. (2011). Gender differences in brain functional connectivity density. *Hum. Brain Mapp.* **33** 849–860.

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

WANG, H. and MARRON, J. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35** 1849–1873. MR2363955

WANG, J., WANG, L., ZANG, Y., YANG, H., TANG, H., GONG, Q., CHEN, Z., ZHU, C. and HE, Y. (2009). Parcellation-dependent small-world brain functional networks: A resting-state fMRI study. *Hum. Brain Mapp.* **30** 1511–1523.

WATSON, G. S. (1983). *Statistics on Spheres*. *University of Arkansas Lecture Notes in the Mathematical Sciences* **6**. Wiley, New York. MR0709262

WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393** 440–442.

XIA, C. (2013). *Eigenvalues in Riemannian Geometry*. IMPA, Rio de Janeiro.

YAN, S., XU, D., ZHANG, B., ZHANG, H.-J., YANG, Q. and LIN, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 40–51.

YAN, C.-G., CRADDOCK, R. C., ZUO, X.-N., ZANG, Y.-F. and MILHAM, M. P. (2013). Standardizing the intrinsic brain: Towards robust measurement of inter-individual variation in 1000 functional connectomes. *NeuroImage* **80** 246–262.

ZIEZOLD, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the* 1974 *European Meeting of Statisticians*.

ZUO, X.-N., EHMKE, R., MENNES, M., IMPERATI, D., CASTELLANOS, F. X., SPORNS, O. and MILHAM, M. P. (2012). Network centrality in the human functional connectome. *Cereb. Cortex* **22** 1862–1875.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 751–770
DOI: 10.1214/16-AOAS1016

ASSIGNMENT OF ENDOGENOUS RETROVIRUS INTEGRATION SITES USING A MIXTURE MODEL

BY DAVID R. HUNTER, LE BAO AND MARY POSS

Pennsylvania State University

Structural variation occurs in the genomes of individuals because of the different positions occupied by repetitive genome elements like endogenous retroviruses, or ERVs. The presence or absence of ERVs can be determined by identifying the junction with the host genome using high-throughput sequence technology and a clustering algorithm. The resulting data give the number of sequence reads assigned to each ERV-host junction sequence for each sampled individual. Variability in the number of reads from an individual integration site makes it difficult to determine whether a site is present for low read counts. We present a novel two-component mixture of negative binomial distributions to model these counts and assign a probability that a given ERV is present in a given individual. We explain how our approach is superior to existing alternatives, including another form of two-component mixture model and the much more common approach of selecting a threshold count for declaring the presence of an ERV. We apply our method to a data set of ERV integrations in mule deer (*Odocoileus hemionus*), a species for which no genomic resources are available, and demonstrate that the discovered patterns of shared integration sites contain information about animal relatedness.

REFERENCES

AKAGI, K., LI, J., STEPHENS, R. M., VOLFOVSKY, N. and SYMER, D. E. (2008). Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res*. **18** 869–880.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans*. *Automat*. *Control* **19** 716–723. MR0423716

BAILLIE, J. K., BARNETT, M. W., UPTON, K. R., GERHARDT, D. J., RICHMOND, T. A., DE SAPIO, F., BRENNAN, P. M., RIZZU, P., SMITH, S., FELL, M., TALBOT, R. T., GUSTINCICH, S., FREEMAN, T. C., MATTICK, J. S., HUME, D. A., HEUTINK, P., CARNINCI, P., JEDDELOH, J. A. and FAULKNER, G. J. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479** 534–537.

BAO, L., ELLEDER, D., MALHOTRA, R., DEGIORGIO, M., MARAVEGIAS, T., HORVATH, L., CARREL, L., GILLIN, C., HRON, T., FÁBRYOVÁ, H., HUNTER, D. R. and POSS, M. (2014). Computational and statistical analyses of insertional polymorphic endogenous retroviruses in a non-model organism. *Comput*. **2** 221–245.

BÖHNE, A., BRUNET, F., GALIANA-ARNOUX, D., SCHULTHEIS, C. and VOLFF, J.-N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res*. **16** 203–15.

Key words and phrases. Mixture model, negative binomial, read count data.

BOURQUE, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Option Genet. Dev.* **19** 607–12.

BRADLEY, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30** 1145–1159.

BURNS, K. H. and BOEKE, J. D. (2012). Human transposon tectonics. *Cell* **149** 740–52.

CONTRERAS-GALINDO, R., KAPLAN, M. H., HE, S., CONTRERAS-GALINDO, A. C., GONZALEZ-HERNANDEZ, M. J., KAPPES, F., DUBE, D., CHAN, S. M., ROBINSON, D., MENG, F., DAI, M., GITLIN, S. D., CHINNAIYAN, A. M., OMENN, G. S. and MARKOVITZ, D. M. (2013). HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Res.* **23** 1505–1513.

CORDAUX, R. and BATZER, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10** 691–703.

CULLINGHAM, C. I., NAKADA, S. M., MERRILL, E. H., BOLLINGER, T. K., PYBUS, M. J. and COLTMAN, D. W. (2011). Multiscale population genetic analysis of mule deer (*Odocoileus hemionus hemionus*) in western Canada sheds new light on the spread of chronic wasting disease. *Can. J. Zool.* **89** 134–147.

ELLEDER, D., KIM, O., PADHI, A., BANKERT, J. G., SIMEONOV, I., SCHUSTER, S. C., WITTEKINDT, N. E., MOTAMENY, S. and POSS, M. (2012). Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *J. Virol.* **86** 2787–96.

EVRONY, G. D., CAI, X., LEE, E., HILLS, L. B., ELHOSARY, P. C., LEHMANN, H. S., PARKER, J. J., ATABAY, K. D., GILMORE, E. C., PODURI, A., PARK, P. J. and WALSH, C. A. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151** 483–496.

EVRONY, G. D., LEE, E., PARK, P. J. and WALSH, C. A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. *eLife* **5** e12966.

FAIRCLOTH, B. C. and GLENN, T. C. (2012). Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS ONE* **7** e42543.

FEDOROFF, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science* **338** 758–767.

HUNTER, D. R., BAO, L. and POSS, M. (2017). Supplement to "Assignment of endogenous retrovirus integration sites using a mixture model." DOI:10.1214/16-AOAS1016SUPP.

ISKOW, R. C., MCCABE, M. T., MILLS, R. E., TORENE, S., PITTARD, W. S., NEUWALD, A. F., VAN MEIR, E. G., VERTINO, P. M. and DEVINE, S. E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141** 1253–61.

KAPUSTA, A., KRONENBERG, Z., LYNCH, V. J., ZHUO, X., RAMSAY, L., BOURQUE, G., YANDELL, M. and FESCHOTTE, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9** e1003470.

KAZAZIAN, H. H. (2004). Mobile elements: Drivers of genome evolution. *Science* **303** 1626–32.

KOKOŠAR, J. and KORDIŠ, D. (2013). Genesis and regulatory wiring of retroelement-derived domesticated genes: A phylogenomic perspective. *Mol. Biol. Evol.* **30** 1015–1031.

LATCH, E. K., REDING, D. M., HEFFELFINGER, J. R., ALCALÁ-GALVÁN, C. H. and RHODES, O. E. (2014). Range-wide analysis of genetic structure in a widespread, highly mobile species (*Odocoileus hemionus*) reveals the importance of historical biogeography. *Mol. Ecol.* **23** 3171–3190.

MALHOTRA, R., ELLEDER, D., BAO, L., HUNTER, D. R., ACHARYA, R. and POSS, M. (2016). Clustering pipeline for determining consensus sequences in targeted next-generation sequencing. In *Proceedings of the* 8*th International Conference on Bioinformatics and Computational Biology* (*BICOB* 2016).

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278.

O'DONNELL, K. A. and BURNS, K. H. (2010). Mobilizing diversity: Transposable element insertions in genetic variation and disease. *Mob. DNA* **1** 21.

POWELL, J. H., KALINOWSKI, S. T., HIGGS, M. D., EBINGER, M. R., VU, N. V. and CROSS, P. C. (2013). Microsatellites indicate minimal barriers to mule deer *Odocoileus hemionus* dispersal across Montana, USA. *Wildl. Biol.* **19** 102–110.

RICHARDSON, S. R., MORELL, S. and FAULKNER, G. J. (2014). L1 retrotransposons and somatic mosaicism in the brain. *Annu. Rev. Genet.* **48** 1–27.

R CORE TEAM (2016). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

WITTEKINDT, N. E., PADHI, A., SCHUSTER, S. C., QI, J., ZHAO, F., TOMSHO, L. P., KASSON, L. R., PACKARD, M., CROSS, P. and POSS, M. (2010). Nodeomics: Pathogen detection in vertebrate lymph nodes using meta-transcriptomics. *PLoS ONE* **5** e13432.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 771–791
DOI: 10.1214/16-AOAS1017

STRUCTURED SUBCOMPOSITION SELECTION IN REGRESSION AND ITS APPLICATION TO MICROBIOME DATA ANALYSIS[1]

BY TAO WANG[*] AND HONGYU ZHAO[*,†]

Shanghai Jiao Tong University[*] and Yale University[†]

Compositional data arise naturally in many practical problems and the analysis of such data presents many statistical challenges, especially in high dimensions. In this article, we consider the problem of subcomposition selection in regression with compositional covariates, where the relationships among the covariates can be represented by a tree with leaf nodes corresponding to covariates. Assuming that the tree structure is available as prior knowledge, we adopt a symmetric version of the linear log contrast model, and propose a tree-guided regularization method for this structured subcomposition selection. Our method is based on a novel penalty function that incorporates the tree structure information node-by-node, encouraging the selection of subcompositions at subtree levels. We show that this optimization problem can be formulated as a generalized lasso problem, the solution of which can be computed efficiently using existing algorithms. An application to a human gut microbiome study and simulations are presented to compare the performance of the proposed method with an l_1 regularization method where the tree structure information is not utilized.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London. MR0865647

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

AKAIKE, H. (1998). *Selected Papers of Hirotugu Akaike*. Springer, New York. MR1486823

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K., GORDON, J. I. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7** 335–336.

CHEN, J., BUSHMAN, F. D., LEWIS, J. D., WU, G. D. and LI, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostat.* **14** 244–258.

FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. MR3065478

FLEISSNER, C. K., HUEBEL, N., ABD EL-BARY, M. M., LOH, G., KLAUS, S. and BLAUT, M. (2010). Absence of intestinal microbiota does not protect mice from diet-induced obesity. *Br. J. Nutr.* **104** 919–929.

Key words and phrases. Compositional data analysis, feature selection, homogeneity, log ratio transformations, penalized regression, phylogenetic tree, the lasso.

GARCIA, T. P., MÜLLER, S., CARROLL, R. J. and WALZEM, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: Application to gut microbiome data. *Bioinformatics* **30** 831–837.

GILL, S. R., POP, M., DEBOY, R. T., ECKBURG, P. B., TURNBAUGH, P. J., SAMUEL, B. S., GORDON, J. I., RELMAN, D. A., FRASER-LIGGETT, C. M. and NELSON, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312** 1355–1359.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

JENATTON, R., AUDIBERT, J.-Y. and BACH, F. (2011). Structured variable selection with sparsity-inducing norms. *J. Mach. Learn. Res.* **12** 2777–2824. MR2854347

JENATTON, R., MAIRAL, J., OBOZINSKI, G. and BACH, F. (2011). Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.* **12** 2297–2334. MR2825428

KIM, S. and XING, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* **5** e1000587.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522

KNIGHTS, D., PARFREY, L. W., ZANEVELD, J., LOZUPONE, C. and KNIGHT, R. (2011). Human-associated microbial signatures: Examining their predictive value. *Cell Host & Microbe* **10** 292–296.

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

LEY, R. E. (2010). Obesity and the human microbiome. *Curr. Opin Gastroenterol.* **26** 5–11.

LI, H. (2015). Microbiome, metagenomics and high-dimensional compositional data analysis. *Ann. Rev. Stat. Appl.* **2** 73–94.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917

MCMURDIE, P. J. and HOLMES, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8** e61217.

NAVAS-MOLINA, J. A., PERALTA-SÁNCHEZ, J. M., GONZÁLEZ, A., MCMURDIE, P. J., VÁZQUEZ-BAEZA, Y., XU, Z., URSELL, L. K., LAUBER, C., ZHOU, H., SONG, S. J., HUNTLEY, J., ACKERMANN, G. L., BERG-LYONS, D., HOLMES, S., CAPORASO, J. G. and KNIGHT, R. (2013). Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* **531** 371–444.

ROTA, G.-C. (1964). The number of partitions of a set. *Amer. Math. Monthly* **71** 498–504. MR0161805

SCEALY, J. L. and WELSH, A. H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 351–375. MR2815780

SCEALY, J. L., DE CARITAT, P., GRUNSKY, E. C., TSAGRIS, M. T. and WELSH, A. H. (2015). Robust principal component analysis for power transformed compositional data. *J. Amer. Statist. Assoc.* **110** 136–148. MR3338492

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10** 1019–1040. MR3528370

ST. JOHN, R. C. (1984). Experiments with mixtures, ill-conditioning, and ridge regression. *J. Qual. Technol.* **16** 81–96.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. and GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444** 1027–1031.

TURNBAUGH, P. J., BÄCKHED, F., FULTON, L. and GORDON, J. I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe* **3** 213–223.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R. et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

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

ZHANG, H., DIBAISE, J. K., ZUCCOLO, A., KUDRNA, D., BRAIDOTTI, M., YU, Y., PARAMESWARAN, P., CROWELL, M. D., WING, R., RITTMANN, B. E. et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc. Natl. Acad. Sci. USA* **106** 2365–2370.

ZHANG, C., ZHANG, M., WANG, S., HAN, R., CAO, Y., HUA, W., MAO, Y., ZHANG, X., PANG, X., WEI, C. et al. (2010). Interactions between gut microbiota, host genetics and diet relevant to development of metabolic syndromes in mice. *ISME J.* **4** 232–241.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

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

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 792–807
DOI: 10.1214/16-AOAS1018

SPATIAL MULTIRESOLUTION ANALYSIS OF THE EFFECT OF $PM_{2.5}$ ON BIRTH WEIGHTS[1]

BY JOSEPH ANTONELLI*, JOEL SCHWARTZ*, ITAI KLOOG† AND BRENT A. COULL*

Harvard Chan School of Public Health and Ben-Gurion University of The Negev†*

Fine particulate matter ($PM_{2.5}$) measured at a given location is a mix of pollution generated locally and pollution traveling long distances in the atmosphere. Therefore, the identification of spatial scales associated with health effects can inform on pollution sources responsible for these effects, resulting in more targeted regulatory policy. Recently, prediction methods that yield high-resolution spatial estimates of $PM_{2.5}$ exposures allow one to evaluate such scale-specific associations. We propose a two-dimensional wavelet decomposition that alleviates restrictive assumptions required for standard wavelet decompositions. Using this method, we decompose daily surfaces of $PM_{2.5}$ to identify which scales of pollution are most associated with adverse health outcomes. A key feature of the approach is that it can remove the purely temporal component of variability in $PM_{2.5}$ levels and calculate effect estimates derived solely from spatial contrasts. This eliminates the potential for unmeasured confounding of the exposure—outcome associations by temporal factors, such as season. We apply our method to a study of birth weights in Massachusetts, U.S.A., from 2003–2008 and find that both local and urban sources of pollution are strongly negatively associated with birth weight. Results also suggest that failure to eliminate temporal confounding in previous analyses attenuated the overall effect estimate toward zero, with the effect estimate growing in magnitude once this source of variability is removed.

REFERENCES

ALEXEEFF, S. E., SCHWARTZ, J., KLOOG, I., CHUDNOVSKY, A., KOUTRAKIS, P. and COULL, B. A. (2015). Consequences of kriging and land use regression for PM2. 5 predictions in epidemiologic analyses: Insights into spatial variability using high-resolution satellite data. *Journal of Exposure Science and Environmental Epidemiology* **25** 138–144.

BREYSSE, P. N., DELFINO, R. J., DOMINICI, F., ELDER, A. C., FRAMPTON, M. W., FROINES, J. R., GEYH, A. S., GODLESKI, J. J., GOLD, D. R., HOPKE, P. K. et al. (2013). US EPA particulate matter research centers: Summary of research results for 2005–2011. *Air Quality, Atmosphere & Health* **6** 333–355.

BROCHU, P. J., KIOUMOURTZOGLOU, M.-A., COULL, B. A., HOPKE, P. K. and SUH, H. H. (2011). Development of a new method to estimate the regional and local contributions to black carbon. *Atmos. Environ.* **45** 7681–7687.

Key words and phrases. Wavelets, multiresolution analysis, spatiotemporal modeling, environmental modeling.

BUHMANN, M. D. (1995). Multiquadric prewavelets on nonequally spaced knots in one dimension. *Math. Comp.* **64** 1611–1625. MR1308448

CHUI, C. K., WARD, J. D., JETTER, K. and STÖCKLER, J. (1996). Wavelets for analyzing scattered data: An unbounded operator approach. *Appl. Comput. Harmon. Anal.* **3** 254–267. MR1400083

DADVAND, P., PARKER, J., BELL, M. L., BONZINI, M., BRAUER, M., DARROW, L. A., GEHRING, U., GLINIANAIA, S. V., GOUVEIA, N., HA, E.-H. et al. (2013). Maternal exposure to particulate air pollution and term birth weight: A multi-country evaluation of effect and heterogeneity.

DARROW, L. A., STRICKLAND, M. J., KLEIN, M., WALLER, L. A., FLANDERS, W. D., CORREA, A., MARCUS, M. and TOLBERT, P. E. (2009). Seasonality of birth and implications for temporal studies of preterm birth. *Epidemiology* **20** 699.

DAUBECHIES, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** 909–996. MR0951745

DOCKERY, D. W., POPE, C. A., XU, X., SPENGLER, J. D., WARE, J. H., FAY, M. E., FERRIS JR, B. G. and SPEIZER, F. E. (1993). An association between air pollution and mortality in six US cities. *N. Engl. J. Med.* **329** 1753–1759.

DOMINICI, F., SHEPPARD, L. and CLYDE, M. (2003). Health effects of air pollution: A statistical review. *Int. Stat. Rev.* **71** 243–276.

DOMINICI, F., PENG, R. D., BELL, M. L., PHAM, L., MCDERMOTT, A., ZEGER, S. L. and SAMET, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *J. Amer. Med. Assoc.* **295** 1127–1134.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GLINIANAIA, S. V., RANKIN, J., BELL, R., PLESS-MULLOLI, T. and HOWEL, D. (2004). Particulate air pollution and fetal health: A systematic review of the epidemiologic evidence. *Epidemiology* **15** 36–45.

GUPTA, C., LAKSHMINARAYAN, C., WANG, S. and MEHTA, A. (2010). Non-dyadic Haar wavelets for streaming and sensor data. In *Data Engineering* (*ICDE*), 2010 *IEEE* 26*th International Conference on* 569–580. IEEE, New York.

HULATA, E., SEGEV, R. and BEN-JACOB, E. (2002). A method for spike sorting and detection based on wavelet packets and Shannon's mutual information. *J. Neurosci. Methods* **117** 1–12.

KLOOG, I., MELLY, S. J., RIDGWAY, W. L., COULL, B. A., SCHWARTZ, J. et al. (2012). Using new satellite based exposure methods to study the association between pregnancy $pm_{2.5}$ exposure, premature birth and birth weight in Massachusetts. *Environ. Health* **11** 1–8.

KLOOG, I., CHUDNOVSKY, A. A., JUST, A. C., NORDIO, F., KOUTRAKIS, P., COULL, B. A., LYAPUSTIN, A., WANG, Y. and SCHWARTZ, J. (2014). A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmos. Environ.* **95** 581–590.

MAYNARD, D., COULL, B. A., GRYPARIS, A. and SCHWARTZ, J. (2007). Mortality risk associated with short-term exposure to traffic particles and sulfates. *Environ. Health Perspect.* 751–755.

MORENO, T., QUEROL, X., ALASTUEY, A., VIANA, M. and GIBBONS, W. (2009). Profiling transient daytime peaks in urban air pollutants: City centre traffic hotspot versus urban background concentrations. *J. Environ. Monit.* **11** 1535–1542.

NENADIC, Z. and BURDICK, J. W. (2005). Spike detection using the continuous wavelet transform. *IEEE Trans. Biomed. Eng.* **52** 74–87.

PENG, R. D., DOMINICI, F. and LOUIS, T. A. (2006). Model choice in time series studies of air pollution and mortality. *J. Roy. Statist. Soc. Ser. A* **169** 179–203. MR2225539

PETROSIAN, A. A. and MEYER, F. G. (2013). *Wavelets in Signal and Image Analysis*: *From Theory to Practice* **19**. Springer Science & Business Media, Dordrecht.

POLLOCK, S. and CASCIO, I. L. (2007). Non-dyadic wavelet analysis. In *Optimisation*, *Econometric and Financial Analysis* 167–203. Springer, Berlin.

POPE III, C. A. (2007). Mortality effects of longer term exposures to fine particulate air pollution: Review of recent epidemiological evidence. *Inhal. Toxicol.* **19** 33–38.

SAMET, J. M., DOMINICI, F., CURRIERO, F. C., COURSAC, I. and ZEGER, S. L. (2000). Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *N. Engl. J. Med.* **343** 1742–1749.

SWELDENS, W. (1998). The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.* **29** 511–546. MR1616507

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TORRENCE, C. and COMPO, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79** 61–78.

WAND, M. P. and ORMEROD, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electron. J. Stat.* **5** 1654–1717. MR2870147

XIONG, R., XU, J. and WU, F. (2006). A lifting-based wavelet transform supporting non-dyadic spatial scalability. In *Image Processing*, 2006 *IEEE International Conference on* 1861–1864. IEEE, New York.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 808–839
DOI: 10.1214/17-AOAS1019

CLUSTERING CORRELATED, SPARSE DATA STREAMS TO ESTIMATE A LOCALIZED HOUSING PRICE INDEX[1]

BY YOU REN, EMILY B. FOX AND ANDREW BRUCE

University of Washington

Understanding how housing values evolve over time is important to policy makers, consumers and real estate professionals. Existing methods for constructing housing indices are computed at a coarse spatial granularity, such as metropolitan regions, which can mask or distort price dynamics apparent in local markets, such as neighborhoods and census tracts. A challenge in moving to estimates at, for example, the census tract level is the scarcity of spatiotemporally localized house sales observations. Our work aims to address this challenge by leveraging observations from multiple census tracts discovered to have correlated valuation dynamics. Our proposed Bayesian nonparametric approach builds on the framework of latent factor models to enable a flexible, data-driven method for inferring the clustering of correlated census tracts. We explore methods for scalability and parallelizability of computations, yielding a housing valuation index at the level of census tract rather than zip code, and on a monthly basis rather than quarterly. Our analysis is provided on a large Seattle metropolitan housing dataset.

REFERENCES

BAILEY, M. J., MUTH, R. F. and NOURSE, H. O. (1963). A Regression Method for Real Estate Price Index Construction. *J. Amer. Statist. Assoc.* **58** 933–942.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

BRUNAUER, W., LANG, S. and UMLAUF, N. (2013). Modelling house prices using multilevel structured additive regression. *Stat. Model.* **13** 95–123. MR3179520

CASE, B. and QUIGLEY, J. M. (1991). The dynamics of real estate prices. *Rev. Econ. Stat.* **73** 50–58.

CASE, K. E. and SHILLER, R. J. (1987). Prices of single family homes since 1970: New indexes for four cities. *N. Engl. Econ. Rev.* 45–56.

CASE, K. E. and SHILLER, R. J. (1989). The efficiency of the market for single-family homes. *Amer. Econ. Rev.* **79** 125–137.

CLEVELAND, R. B., CLEVELAND, W. S., MCRAE, J. E. and TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on loess (with discussion). *J. Off. Stat.* **6** 3–73.

ENGLUND, P., QUIGLEY, J. M. and REDFEARN, C. L. (1999). The choice of methodology for computing housing price indexes: Comparisons of temporal aggregation and sample definition. *J. Real Estate Finance Econ.* **19** 91–112.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

Key words and phrases. Bayesian nonparametrics, clustering, housing price index, multiple time series, state space models.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

GATZLAFF, D. H. and HAURIN, D. R. (1997). Sample selection bias and repeat-sales index estimates. *J. Real Estate Finance Econ.* **14** 33–50.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GOETZMANN, W. N. and PENG, L. (2002). The bias of the RSR estimator and the accuracy of some alternatives. *Real Estate Econ.* **30** 13–39.

IACOVIELLO, M. (2011). Housing wealth and consumption. Board of Governors of the Federal Reserve System, International Finance Discussion Papers 1027.

LIAO, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognit.* **38** 1857–1874.

MACLAURIN, D. and ADAMS, R. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Uncertainty in Artificial Intelligence*.

MEESE, R. A. and WALLACE, N. E. (1997). The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression, and hybrid approaches. *J. Real Estate Finance Econ.* **14** 51–73.

MUNKRES, J. (1957). Algorithms for the assignment and transportation problems. *J. Soc. Indust. Appl. Math.* **5** 32–38. MR0093429

NAGARAJA, C. H., BROWN, L. D. and ZHAO, L. H. (2011). An autoregressive approach to house price modeling. *Ann. Appl. Stat.* **5** 124–149. MR2810392

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

NIETO-BARAJAS, L. E. and CONTRERAS-CRISTÁN, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Anal.* **9** 147–169. MR3188303

PALLA, K., GHAHRAMANI, Z. and KNOWLES, D. (2012). A nonparametric variable clustering model. *Adv. Neural Inf. Process. Syst.* **25** 2987–2995.

PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

REN, Y., FOX, E. B. and BRUCE, A. (2017). Supplement to "Clustering correlated, sparse data streams to estimate a localized housing price index." DOI:10.1214/17-AOAS1019SUPP.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHILLER, R. (1991). Arithmetic repeat sales price estimators. *J. Housing Econ.* **1** 110–126.

SMITH, P. L. (1979). Splines as a useful and convenient statistical tool. *Amer. Statist.* **33** 57–62.

WILLIAMSON, S., DUBEY, A. and XING, E. (2013). Parallel Markov chain Monte Carlo for nonparametric mixture models. In *International Conference on Machine Learning* 98–106.

ZILLOW (2014). Zillow home value index: Methodology. http://www.zillow.com/research/zhvi-methodology-6032/.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 840–867
DOI: 10.1214/17-AOAS1020

NONPARAMETRIC ESTIMATION OF PREGNANCY OUTCOME PROBABILITIES[1]

BY SARAH FRIEDRICH[*], JAN BEYERSMANN[*], URSULA WINTERFELD[†], MARTIN SCHUMACHER[‡] AND ARTHUR ALLIGNOL[*]

Ulm University[*], University Hospital Lausanne[†] and University Medical Center Freiburg[‡]

Estimating pregnancy outcome probabilities based on observational cohorts has to account for both left-truncation, because the time scale is gestational age, and for competing risks, because, for example, an induced abortion may be precluded by a spontaneous abortion. The applied aim of this work was to investigate the impact of statins on pregnancy outcomes using data from Teratology Information Services. Using the standard Aalen–Johansen estimator of the cumulative event probabilities suggested the medically unexpected finding that statin exposure decreased the probability of induced abortion and led to more live births. The reason was an early induced abortion in a very small risk set in the control group, leading to unstable estimation which propagated over the whole time span. We suggest a stabilized Aalen–Johansen estimator which discards contributions from overly small risk sets. The decision whether a risk set is considered overly small is controlled by tuning parameters which we choose using a cross-validated Brier Score. We show that the new estimator enjoys the same asymptotic properties as the original Aalen–Johansen estimator. Small sample properties are investigated in extensive simulations. We also discuss extensions to more general multistate models.

REFERENCES

AALEN, O. O. and JOHANSEN, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Stat.* **5** 141–150. MR0509450

ALLIGNOL, A., SCHUMACHER, M. and BEYERSMANN, J. (2010). A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biom. J.* **52** 126–137. MR2756598

ANDERSEN, P. K. and KEIDING, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Stat. Med.* **31** 1074–1088. MR2925679

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

ANDERSEN, A.-M. N., ANDERSEN, P. K., OLSEN, J., GRØNBÆK, M. and STRANDBERG-LARSEN, K. (2012). Moderate alcohol intake during pregnancy and risk of fetal death. *Int. J. Epidemiol.* **41** 405–413.

BEYERSMANN, J., ALLIGNOL, A. and SCHUMACHER, M. (2012). *Competing Risks and Multistate Models with R*. Springer, New York. MR3025354

BEYERSMANN, J. and SCHUMACHER, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics* **9** 765–776.

Key words and phrases. Left-truncation, pregnancy, competing risks, abortion, Brier score.

BEYERSMANN, J., LATOUCHE, A., BUCHHOLZ, A. and SCHUMACHER, M. (2009). Simulating competing risks data in survival analysis. *Stat. Med.* **28** 956–971. MR2518359

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331. MR0711106

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560. MR1467848

FRIEDRICH, S., BEYERSMANN, J., WINTERFELD, U., SCHUMACHER, M. and ALLIGNOL, A. (2017). Supplement to "Nonparametric estimation of pregnancy outcome probabilities." DOI:10.1214/17-AOAS1020SUPP.

GERDS, T. A. and SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** 1029–1040. MR2312613

GERDS, T. A. and SCHUMACHER, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63** 1283–1287, 1316. MR2414608

GOOLEY, T. A., LEISENRING, W., CROWLEY, J., STORER, B. E. et al. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Stat. Med.* **18** 695–706.

GRZESKOWIAK, L. E., GILBERT, A. L. and MORRISON, J. L. (2012). Exposed or not exposed? Exploring exposure classification in studies using administrative data to investigate outcomes following medication use during pregnancy. *Eur. J. Clin. Pharmacol.* **68** 459–67.

HANCOCK, R. L., KOREN, G., EINARSON, A. and UNGAR, W. J. (2007). The effectiveness of teratology information services (TIS). *Reprod. Toxicol.* (*Elmsford N.Y.*) **23** 125–132.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HELD, L. and SABANÉS BOVÉ, D. (2014). *Applied Statistical Inference*. Springer, Heidelberg. MR3155114

KEIDING, N. and GILL, R. D. (1990). Random truncation models and Markov processes. *Ann. Statist.* **18** 582–602. MR1056328

KLEIN, J. P. (1991). Small sample moments of some estimators of the variance of the Kaplan–Meier and Nelson–Aalen estimators. *Scand. J. Stat.* **18** 333–340. MR1157787

LAI, T. L. and YING, Z. (1991). Estimating a distribution function with truncated and censored data. *Ann. Statist.* **19** 417–442. MR1091860

LUPATTELLI, A., SPIGSET, O., TWIGG, M. J. et al. (2014). Medication use in pregnancy: A cross-sectional, multinational web-based study. *BMJ Open* **4**.

MACKENZIE, T. (2012). Survival curve estimation with dependent left truncated data using Cox's model. *Int. J. Biostat.* **8** Art. 29. MR2997679

MEISTER, R. and SCHAEFER, C. (2008). Statistical methods for estimating the probability of spontaneous abortion in observational studies—analyzing pregnancies exposed to coumarin derivatives. *Reprod. Toxicol.* (*Elmsford N.Y.*) **26** 31–35.

SCHOOP, R., BEYERSMANN, J., SCHUMACHER, M. and BINDER, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biom. J.* **53** 88–112. MR2767380

SLAMA, R., BALLESTER, F., CASAS, M., CORDIER, S., EGGESBØ, M., INIGUEZ, C., NIEUWENHUIJSEN, M., PHILIPPAT, C., REY, S., VANDENTORREN, S. et al. (2014). Epidemiologic tools to study the influence of environmental factors on fecundity and pregnancy-related outcomes. *Epidemiol. Rev.* **36** 148–164.

TSAI, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77** 169–177. MR1049418

VAN HOUWELINGEN, H. C. and PUTTER, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. *Monographs on Statistics and Applied Probability* **123**. CRC Press, Boca Raton, FL. MR3058205

WILLAND, I. (2011). *Statistisches Jahrbuch*. Statistisches Bundesamt, Wiesbaden.

WILLAND, I. (2014). *Statistisches Jahrbuch*. Statistisches Bundesamt, Wiesbaden.

WINTERFELD, U., ALLIGNOL, A., PANCHAUD, A., ROTHUIZEN, L. E., MERLOB, P., CUPPERS-MAARSCHALKERWEERD, B., VIAL, T., STEPHENS, S., CLEMENTI, M., SANTIS, M. et al. (2013). Pregnancy outcome following maternal exposure to statins: A multicentre prospective study. *BJOG*: *An International Journal of Obstetrics and Gynaecology* **120** 463–471.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 868–897
DOI: 10.1214/17-AOAS1021

BAYESIAN INFERENCE OF HIGH-DIMENSIONAL, CLUSTER-STRUCTURED ORDINARY DIFFERENTIAL EQUATION MODELS WITH APPLICATIONS TO BRAIN CONNECTIVITY STUDIES

BY TINGTING ZHANG[*,1,2], QIANNAN YIN[*], BRIAN CAFFO[†], YINGE SUN[*] AND DANA BOATMAN-REICH[†,1,3]

*University of Virginia * and Johns Hopkins University†*

We build a new ordinary differential equation (ODE) model for the directional interaction, also called effective connectivity, among brain regions whose activities are measured by intracranial electrocorticography (ECoG) data. In contrast to existing ODE models that focus on effective connectivity among only a few large anatomic brain regions and that rely on strong prior belief of the existence and strength of the connectivity, the proposed high-dimensional ODE model, motivated by statistical considerations, can be used to explore connectivity among multiple small brain regions. The new model, called the modular and indicator-based dynamic directional model (MIDDM), features a cluster structure, which consists of modules of densely connected brain regions, and uses indicators to differentiate significant and void directional interactions among brain regions. We develop a unified Bayesian framework to quantify uncertainty in the assumed ODE model, identify clusters, select strongly connected brain regions, and make statistical comparison between brain networks across different experimental trials. The prior distributions in the Bayesian model for MIDDM parameters are carefully designed such that the ensuing joint posterior distributions for ODE state functions and the MIDDM parameters have well-defined and easy-to-simulate posterior conditional distributions. To further speed up the posterior simulation, we employ parallel computing schemes in Markov chain Monte Carlo steps. We show that the proposed Bayesian approach outperforms an existing optimization-based ODE estimation method. We apply the proposed method to an auditory electrocorticography dataset and evaluate brain auditory network changes across trials and different auditory stimuli.

REFERENCES

AERTSEN, A. and PREISSL, H. (1991). Dynamics of activity and connectivity in physiological neuronal networks. In *Nonlinear Dynamics and Neuronal Networks* (H. Schuster, ed.) 281–302. VCH publishers Inc, New York.

ANDERSON, J. (2005). Learning in sparsely connected and sparsely coded system. Ersatz Brain Project Working Note.

BARD, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York. MR0326870

BHAUMIK, P. and GHOSAL, S. (2014). Bayesian estimation in differential equation models. Preprint. Available at arXiv:1403.0609.

Key words and phrases. Bayesian inference, ODE models, cluster structure, directional brain networks, network edge selection.

BIEGLER, L., DAMIANO, J. and BLAU, G. (1986). Nonlinear parameter estimation: A case study comparison. *AIChE J.* **32** 29–45.

BOATMAN-REICH, D., FRANASZCZUK, P. J., KORZENIEWSKA, A., CAFFO, B., RITZL, E. K., COLWELL, S. and CRONE, N. E. (2010). Quantifying auditory event-related responses in multichannel human intracranial recordings. *Front. Comput. Neurosci.* **4** 4.

BRESSLER, S. and DING, M. (2002). Event-related potentials. In *The Handbook of Brain Theory and Neural Networks* 412–415. Wiley, New York.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. MR1626005

BRUNEL, N. J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electron. J. Stat.* **2** 1242–1267. MR2471285

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198.

CAFFO, B., PENG, R., DOMINICI, F., LOUIS, T. A. and ZEGER, S. (2011). Parallel MCMC for analyzing distributed lag models with systematic missing data for an application in environmental epidemiology. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X. Meng, eds.) 493–511. CRC Press, Boca Raton, FL.

CALDERHEAD, B., GIROLAMI, M. and LAWRENCE, N. (2008). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Adv. Neural Inf. Process. Syst.* **22**.

CAMPBELL, D. A. (2007). *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–McGill University (Canada). MR2711737

CAO, J., HUANG, J. Z. and WU, H. (2012). Penalized nonlinear least squares estimation of time-varying parameters in ordinary differential equations. *J. Comput. Graph. Statist.* **21** 42–56. MR2913355

CATANI, M., DELLACQUA, F., VERGANI, F., MALIK, F., HODGE, H., ROY, P., VALABREGUE, R. and THIEBAUT DE SCHOTTEN, M. (2012). Short frontal lobe connections of the human brain. *Cortex* **48** 273–291.

CERVENKA, M. C., FRANASZCZUK, P. J., CRONE, N. E., HONG, B., CAFFO, B. S., BHATT, P., LENZ, F. A. and BOATMAN-REICH, D. (2013). Reliability of early cortical auditory gamma-band responses. *Clin. Neurophysiol.* **124** 70–82.

CHEUNG, S., OLIVER, T., PRUDENCIO, E., PRUDHOMME, S. and MOSER, R. (2011). Bayesian uncertainty analysis with applications to turbulence modeling. *Reliab. Eng. Syst. Saf.* **96** 1137–1149.

CHKREBTII, O. A., CAMPBELL, D. A., CALDERHEAD, B. and GIROLAMI, M. A. (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Anal.* **11** 1239–1267. MR3577378

CONRAD, P., GIROLAMI, M., SÄRKKÄ, S., STUART, A. and ZYGALAKIS, K. (2015). Probability Measures for Numerical Solutions of Differential Equations.

DAUNIZEAU, J., DAVID, O. and STEPHAN, K. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage* **58** 312–322.

DAVID, O. and FRISTON, K. (2003). A neural mass model for MEG/EEG: Coupling and neuronal dynamics. *NeuroImage* **20** 1743–1755.

DAVID, O., KIEBEL, S., HARRISON, L., MATTOUT, J., KILNER, J. and FRISTON, K. (2006). Dynamic causal modelling of evoked responses in EEG and MEG. *NeuroImage* **30** 1255–1272.

DEUFLHARD, P. and BORNEMANN, F. (2002). *Scientific Computing with Ordinary Differential Equations*. Springer, New York. MR1912409

DUNSON, D. B., HERRING, A. H. and ENGEL, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.* **103** 534–546. MR2523991

DURKA, P., IRCHA, D., NEUPER, C. and PFURTSCHELLER, G. (2001). Time-frequency microstructure of event-related electro-encephalogram eesynchronisation and synchronisation. *Med. Biol. Eng. Comput.* **39** 315–3211.

ELIADES, S., CRONE, N., ANDERSON, W., RAMADOSS, D., LENZ, F. and BOATMAN-REICH, D. (2014). Adaptation of high-gamma responses in human auditory association cortex. *J. Neurophysiol.* **112** 2147–2163.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FÖLDIÁK, P. and YOUNG, M. P. (1995). Sparse coding in the primate cortex. In *The Handbook of Brain Theory and Neural Networks* 895–898. MIT Press, Cambridge.

FRANASZCZUK, P. J. and BERGEY, G. K. (1998). Application of the directed transfer function method to mesial and lateral onset temporal lobe seizures. *Brain Topogr.* **11** 13–21.

FRISTON, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biology* **7** 33.

FRISTON, K., HARRISON, L. and PENNY, W. (2003). Dynamic causal modelling. *NeuroImage* **19** 1273–1302.

GARRIDO, M., KILNER, J., KIEBEL, S., STEPHAN, K., BALDEWEG, T. and FRISTON, K. (2009). Comparative frequency analysis of single EEG-evoked potential records. *NeuroImage* **48** 269–279.

GELMAN, A., BOIS, F. and JIANG, J. (1996). Physiological pharamacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Statist. Assoc.* **91** 1400–1412.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. and MCCULLOCH, R. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GIROLAMI, M. (2008). Bayesian inference for differential equations. *Theoret. Comput. Sci.* **408** 4–16. MR2460604

GRANER, F. and GLAZIER, J. A. (1992). Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys. Rev. Lett.* **69** 2013–2016.

HEMKER, P. (1972). Numerical methods for differential equations in system simulations and in parameter estimation. *Analysis and Simulation of Biochemical Systems* 59–80.

HERRMANN, B., HENRY, M. and OBLESER, J. (2013). Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. *J. Neurophysiol.* **109** 2086–2096.

HERRMANN, B., SCHLICHTING, N. and OBLESER, J. (2014). Dynamic range adaptation to spectral stimulus statistics in human auditory cortex. *J. Neurosci.* **34** 327–331.

HUANG, Y., LIU, D. and WU, H. (2006). Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62** 413–423. MR2227489

HUANG, Y. and WU, H. (2006). A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *J. Appl. Stat.* **33** 155–174. MR2223142

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398

KIEBEL, S., DAVID, O. and FRISTON, K. (2006). Dynamic causal modelling of evoked responses in EEG/MEG with lead-field parameterization. *NeuroImage* **30** 1273–1284.

KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877–893. MR2285077

LI, Z., OSBORNE, M. R. and PRVAN, T. (2005). Parameter estimation of ordinary differential equations. *IMA J. Numer. Anal.* **25** 264–285. MR2126204

LU, T., LIANG, H., LI, H. and WU, H. (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *J. Amer. Statist. Assoc.* **106** 1242–1258. MR2896833

MATTHEIJ, R. and MOLENAAR, J. (2002). *Ordinary Differential Equations in Theory and Practice. Classics in Applied Mathematics* **43**. SIAM, Philadelphia, PA. MR1946758

MICHELOYANNIS, S. (2012). Graph-based network analysis in schizophrenia. *World J. Psychiatry* **2** 1–12.

MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2001193

MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

MILO, R., ITZKOVITZ, S., KASHTAN, N., LEVITT, R., SHEN-ORR, S., AYZENSHTAT, I., SHEFFER, M. and ALON, U. (2004). Superfamilies of evolved and designed networks. *Science* **303** 1538–1542.

NÄÄTÄNEN, R., PAAVILAINEN, P., RINNE, T. and ALHO, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clin. Neurophysiol.* **118** 2544–2590.

NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8696.

OLIVER, T. and MOSER, R. (2011). Bayesian uncertainty quantification applied to RANS turbulence models. *Int. J. Mod. Phys. Conf. Ser.* **318** 042032.

OLSHAUSEN, B. and FIELD, D. (2004). Sparse coding of sensor inputs. *Current Opinions in Neurobiology* **14** 481–487.

PARK, H.-J. and FRISTON, K. (2013). Structural and functional brain networks: From connections to cognition. *Science* **342** 1238411.

POTTS, R. B. (1952). Some generalized order-disorder transformations. *Math. Proc. Cambridge Philos. Soc.* **48** 106–109. MR0047571

POYTON, A., VARZIRI, M., MCAULEY, K., MCLELLAN, P. and RAMSAY, J. (2006). Parameter estimation in continuous dynamic models using principal differential analysis. *Computational Chemical Engineering* **30** 698–708.

QI, X. and ZHAO, H. (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Ann. Statist.* **38** 435–481. MR2589327

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 741–796. MR2368570

REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. MR2411660

REISS, P. T. and OGDEN, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 505–523. MR2649608

SCHÖNWIESNER, M., NOVITSKI, N., PAKARINEN, S., CARLSON, S., TERVANIEMI, M. and NÄÄTÄNEN, R. (2007). Heschl's gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. *J. Neurophysiol.* **97** 2075–2082.

SINAI, A., CRONE, N., WIED, H., FRANASZCZUK, P., MIGLIORETTI, D. and BOATMAN-REICH, D. (2009). Intracranial mapping of auditory perception: Event-related responses and electrocortical stimulation. *Clin. Neurophysiol.* **120** 140–149.

SPORNS, O. (2011). *Networks of the Brain*. MIT Press, Cambridge, MA.
STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785
TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563
THEO, H. and MIKE, E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol* **36** 261–279.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103** 790–796. MR2524010
VARAH, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.* **3** 28–46. MR0651865
VOIT, E. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge Univ. Press, Cambridge.
WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA. MR1045442
WANG, H. and LENG, C. (2008). A note on adaptive group lasso. *Comput. Statist. Data Anal.* **52** 5277–5286. MR2526593
WU, H., LU, T., XUE, H. and LIANG, H. (2014a). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *J. Amer. Statist. Assoc.* **109** 700–716. MR3223744
WU, S., XUE, H., WU, Y. and WU, H. (2014b). Variable selection for sparse high-dimensional nonlinear regression models by combining nonnegative garrote and sure independence screening. *Statist. Sinica* **24** 1365–1387. MR3241292
XUE, H., MIAO, H. and WU, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.* **38** 2351–2387. MR2676892
YI, N., GEORGE, V. and ALLISON, D. B. (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164** 1129–1138.
YUAN, M. and LIN, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100** 1215–1225. MR2236436
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
ZHANG, T., WU, J., LI, F., CAFFO, B. and BOATMAN-REICH, D. (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *J. Amer. Statist. Assoc.* **110** 93–106. MR3338489
ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469
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

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 898–918
DOI: 10.1214/17-AOAS1022

EVALUATION OF THE COOLING TREND IN THE IONOSPHERE USING FUNCTIONAL REGRESSION WITH INCOMPLETE CURVES[1]

BY OLEKSANDR GROMENKO, PIOTR KOKOSZKA AND JAN SOJKA

IBM Research, Colorado State University and Utah State University

We develop a statistical framework to test the hypothesis of the existence of an ionospheric cooling trend related to the global warming hypothesis; both are attributed to the same driver, namely the increased concentration of greenhouse gases. However, the study of a temporal trend in the ionosphere is easier because there are fewer covariates to be taken into account. The hypothesis that a cooling trend in the ionosphere exists has been an important focus of space physics research for over two decades. A central difficulty in reaching broadly agreed—on conclusions has been the absence of data with sufficiently long temporal and sufficiently broad spatial coverage. Complete time series of data that cover several decades exist only in a few separated (industrialized) regions. The space physics community has struggled to combine the information contained in these data, and often contradictory conclusions have been reported based on the analyses relying on one or a few locations. We present a statistical analysis that uses all data, even those with incomplete temporal coverage. It is based on a new functional regression approach that can handle spatially indexed curves whose temporal domain depends on location and may contain gaps. The test statistic combines spatial and temporal dependence in the data and is approximately normally distributed. We conclude that a statistically significant cooling trend exists in the Northern Hemisphere. This confirms the hypothesis put forward in the space physics community over two decades ago.

REFERENCES

BAKAR, K. and SAHU, S. (2015). spTimer: Spatio-temporal Bayesian modeling using R. *J. Stat. Softw.* **63** 1–32.

BREMER, J., DAMBOLDT, T., MIELICH, J. and SUESSMANN, P. (2012). Comparing long-term trends in the ionospheric F2—Region with two different methods. *Journal of Atmospheric and Solar-Terrestrial Physics* **77** 174–185.

CRAINICEANU, C. M., STAICU, A.-M., RAY, S. and PUNJABI, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Stat. Med.* **31** 3223–3240. MR2993623

CRESSIE, N. and HAWKINS, D. M. (1980). Robust estimation of the variogram. I. *J. Int. Assoc. Math. Geol.* **12** 115–125. MR0595404

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ. MR2848400

Key words and phrases. Cooling trend, functional regression, incomplete time series, ionosphere, solar activity, spatial averaging, spatio-temporal modeling.

DAMBOLDT, T. and SUESSMANN, P. (2012). Consolidated Database of worldwide measured monthly medians of ionospheric characteristics foF2 and M(3000)F2. INAG Bulletin on the Web, INAG-73. Available at www.ips.gov.au/IPSHosted/INAG/web-73/index.html.

DELICADO, P., GIRALDO, R., COMAS, C. and MATEU, J. (2010). Statistics for spatial functional data: Some recent contributions. *Environmetrics* **21** 224–239. MR2842240

GELFAND, A. E., DIGGLE, P., GUTTORP, P. and FUENTES, M., eds. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

GENTON, M. G. (1998). Highly robust variogram estimation. *Math. Geol.* **30** 213–221. MR1610687

GENTON, M. G. (2007). Separable approximations of space–time covariance matrices. *Environmetrics* **18** 681–695. MR2408938

GIRALDO, R., DELICADO, P. and MATEU, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *J. Agric. Biol. Environ. Stat.* **15** 66–82. MR2755385

GIRALDO, R., DELICADO, P. and MATEU, J. (2011). Ordinary kriging for function-valued spatial data. *Environ. Ecol. Stat.* **18** 411–426. MR2832903

GIRALDO, R., DELICADO, P. and MATEU, J. (2012). Hierarchical clustering of spatially correlated functional data. *Stat. Neerl.* **66** 403–421. MR2983302

GROMENKO, O. and KOKOSZKA, P. (2012). Testing the equality of mean functions of ionospheric critical frequency curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 715–731. MR2993506

GROMENKO, O. and KOKOSZKA, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Comput. Statist. Data Anal.* **59** 82–94. MR3000043

GROMENKO, O., KOKOSZKA, P., ZHU, L. and SOJKA, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Ann. Appl. Stat.* **6** 669–696. MR2976487

HAAS, T. C. (1995). Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Statist. Assoc.* **90** 1189–1199.

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer, New York. MR2920735

JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427

KELLY, M. C. (2009). *The Earth's Ionosphere*, 2nd ed. Academic Press, San Diego, CA.

KRAUS, D. (2015). Components and completion of partially observed functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 777–801. MR3382597

LASTOVICKA, J., SOLOMON, S. C. and QIAN, L. (2012). Trends in the neutral and ionized atmosphere. *Reviews of Space Physics* **168** 113–145.

LASTOVICKA, J., MIKHAILOV, A. V., ULICH, T., BREMER, J., ELIAS, A., ORTIZ DE ADLER, N., JARA, V., ABBARCA DEL RIO, R., FOPPIANO, A., OVALLE, E. and DANILOV, A. (2006). Long term trends in foF2: A comparison of various methods. *Journal of Atmospheric and Solar-Terrestrial Physics* **68** 1854–1870.

LIEBL, D. (2013). Modeling and forecasting electricity spot prices: A functional data perspective. *Ann. Appl. Stat.* **7** 1562–1592. MR3127959

LUTTINEN, J. and ILIN, A. (2009). Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems* 22 (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1177–1185. Curran Associates, Red Hook, NY.

LUTTINEN, J. and ILIN, A. (2012). Efficient Gaussian process inference for short-scale spatio–temporal modeling. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* **22**. JMLR W&CP.

MIELICH, J. and BREMER, J. (2013). Long-term trends in the ionospheric F2 region with different solar activity indices. *Annals of Geophysics* **31** 291–303.

NERINI, D., MONESTIEZ, P. and MANTÉ, C. (2010). Cokriging for spatial functional data. *J. Multivariate Anal.* **101** 409–418. MR2564350

PAUL, D. and PENG, J. (2011). Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach. *Electron. J. Stat.* **5** 1960–2003. MR2870154

RISHBETH, H. (1990). A greenhouse effect in the ionosphere? *Planetary and Space Science* **38** 945–948.

ROBLE, R. G. and DICKINSON, R. E. (1989). How will changes in carbon dioxide and methane modify the mean structure of the mesosphere and thermosphere? *Geophysical Research Letters* **16** 1441–1444.

SECCHI, P., VANTINI, S. and VITELLI, V. (2011). A clustering algorithm for spatially dependent functional data. *Procedia Environmental Sciences* **7** 176–181.

SECCHI, P., VANTINI, S. and VITELLI, V. (2012). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation* **22** 53–64.

SHERMAN, M. (2011). *Spatial Statistics and Spatio-Temporal Data*: *Covariance Functions and Directional Properties*. Wiley, Chichester. MR2815783

STAICU, A.-M., CRAINICEANU, C. and CARROLL, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics* **11** 177–194.

STAICU, A.-M., CRAINICEANU, C. M., REICH, D. S. and RUPPERT, D. (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics* **68** 331–343. MR2959599

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. Springer, New York. MR1697409

STEIN, M. L. (2005). Space–time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. MR2156840

SUN, Y., LI, B. and GENTON, M. G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space–Time Modelling of Natural Events* (E. Porcu, J. M. Montero and M. Schlather, eds.) **3** 55–77. Springer, Berlin.

THÉBAULT, E., FINLAY, C. C., BEGGAN, C. D., ALKEN, P., AUBERT, J., BARROIS, O., BERTRAND, F., BONDAR, T., BONESS, A., BROCCO, L., CANET, E., CHAMBODUT, A., CHULLIAT, A., COÏSSON, P., CIVET, F., DU, A., FOURNIER, A., FRATTER, I., GILLET, N., HAMILTON, B., HAMOUDI, M., HULOT, G., JAGER, T., KORTE, M., KUANG, W., LALANNE, X., LANGLAIS, B., LÉGER, J.-M., LESUR, V., LOWES, F. J., MACMILLAN, S., MANDEA, M., MANOJ, C., MAUS, S., OLSEN, N., PETROV, V., RIDLEY, V., ROTHER, M., SABAKA, T. J., SATURNINO, D., SCHACHTSCHNEIDER, R., SIROL, O., TANGBORN, A., THOMSON, A., TØFFNER-CLAUSEN, L., VIGNERON, P., WARDINSKI, I. and ZVEREVA, T. (2015). International geomagnetic reference field: The 12th generation. *Earth, Planets and Space* **67** 1–19.

ULICH, T., CLILVERD, M. A. and RISHBETH, H. (2003). Determining long-term change in the ionosphere. *Eos, Transactions American Geophysical Union* **84** 581–585.

YANG, J., ZHU, H., CHOI, T. and COX, D. D. (2016). Smoothing and mean-covariance estimation of functional data with a Bayesian hierarchical model. *Bayesian Anal.* **11** 649–670. MR3498041

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 919–948
DOI: 10.1214/17-AOAS1023

VACCINES, CONTAGION, AND SOCIAL NETWORKS

BY ELIZABETH L. OGBURN[1] AND TYLER J. VANDERWEELE[2]

Johns Hopkins School of Public Health and Harvard School of Public Health

Consider the causal effect that one individual's treatment may have on another individual's outcome when the outcome is contagious, with specific application to the effect of vaccination on an infectious disease outcome. The effect of one individual's vaccination on another's outcome can be decomposed into two different causal effects, called the "infectiousness" and "contagion" effects. We present identifying assumptions and estimation or testing procedures for infectiousness and contagion effects in two different settings: (1) using data sampled from independent groups of observations, and (2) using data collected from a single interdependent social network. The methods that we propose for social network data require fitting generalized linear models (GLMs). GLMs and other statistical models that require independence across subjects have been used widely to estimate causal effects in social network data, but because the subjects in networks are presumably not independent, the use of such models is generally invalid, resulting in inference that is expected to be anticonservative. We describe a subsampling scheme that ensures that GLM errors are uncorrelated across subjects despite the fact that outcomes are nonindependent. This simultaneously demonstrates the possibility of using GLMs and related statistical models for network data and highlights their limitations.

REFERENCES

ALI, M. M. and DWYER, D. S. (2009). Estimating peer effects in adolescent smoking behavior: A longitudinal analysis. *J. Adolesc. Health* **45** 402–408.

ANDERSON, R. M. and MAY, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature* **318** 323–329.

ARONOW, P. M. and SAMII, C. (2013). Estimating average causal effects under general interference. Technical report, arXiv:1305.6156.

BOWERS, J., FREDRICKSON, M. M. and PANAGOPOULOS, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* **21** 97–124.

BRESLOW, N. E. (1996). Generalized linear models: Checking assumptions and strengthening conclusions. *Stat. Appl.* **8** 23–41.

CACIOPPO, J. T., FOWLER, J. H. and CHRISTAKIS, N. A. (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *J. Pers. Soc. Psychol.* **97** 977.

CHOI, D. S. (2014). Estimation of monotone treatment effects in network experiments. ArXiv Preprint, arXiv:1408.4102.

CHRISTAKIS, N. A. and FOWLER, J. H. (2007). The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357** 370–379.

CHRISTAKIS, N. A. and FOWLER, J. H. (2008). The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358** 2249–2258.

Key words and phrases. Causal inference, social networks, contagion, peer effects.

CHRISTAKIS, N. A. and FOWLER, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5** e12948.

CHRISTAKIS, N. A. and FOWLER, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Stat. Med.* **32** 556–577. MR3042499

COHEN-COLE, E. and FLETCHER, J. M. (2008). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *J. Health Econ.* **27** 1382–1387.

EAMES, K. T. D. and KEELING, M. J. (2002). Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Natl. Acad. Sci. USA* **99** 13330–13335.

EAMES, K. T. D. and KEELING, M. J. (2004). Monogamous networks and the spread of sexually transmitted diseases. *Math. Biosci.* **189** 115–130. MR2065412

EARN, D. J. D., DUSHOFF, J. and LEVIN, S. A. (2002). Ecology and evolution of the flu. *Trends Ecol. Evol.* **17** 334–340.

ECKLES, D., KARRER, B. and UGANDER, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. ArXiv Preprint, arXiv:1404.7530.

EUBANK, S., GUCLU, H., KUMAR, V. S. A., MARATHE, M. V., SRINIVASAN, A., TOROCZKAI, Z. and WANG, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature* **429** 180–184.

FINE, P. E. M. (1993). Herd immunity: History, theory, practice. *Epidemiol. Rev.* **15** 265–302.

FOWLER, J. H. and CHRISTAKIS, N. A. (2008). Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher; and Trogdon, Nonnemaker, and Pais. *J. Health Econ.* **27** 1400–1405.

GILL, J. (2001). *Generalized Linear Models*: *A Unified Approach*. *Quantitative Applications in the Social Sciences* **134**. Sage Publications, Thousand Oaks, CA.

HALLORAN, M. E. and HUDGENS, M. G. (2012). Causal inference for vaccine effects on infectiousness. *Int. J. Biostat.* **8** Art. 6. MR2925328

HALLORAN, M. E. and STRUCHINER, C. J. (1991). Study designs for dependent happenings. *Epidemiology* **2** 331–338.

HALLORAN, M. E. and STRUCHINER, C. J. (1992). Modeling transmission dynamics of stage-specific malaria vaccines. *Parasitol. Today* (*Regul. Ed.*) **8** 77–85.

HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6** 142–151.

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334.

JOHN, T. J. and SAMUEL, R. (2000). Herd immunity and herd effect: New insights and definitions. *Eur. J. Epidemiol.* **16** 601–606.

KEELING, M. J. and EAMES, K. T. (2005). Networks and epidemic models. *J. R. Soc. Interface* **2** 295–307.

KELLER, M. A. and STIEHM, E. R. (2000). Passive immunity in prevention and treatment of infectious diseases. *Clin. Microbiol. Rev.* **13** 602–614.

KLOVDAHL, A. S. (1985). Social networks and the spread of infectious diseases: The AIDS example. *Soc. Sci. Med.* **21** 1203–1216.

KLOVDAHL, A. S., POTTERAT, J. J., WOODHOUSE, D. E., MUTH, J. B., MUTH, S. Q. and DARROW, W. W. (1994). Social networks and infectious disease: The Colorado Springs study. *Soc. Sci. Med.* **38** 79–88.

LATORA, V., NYAMBA, A., SIMPORE, J., SYLVETTE, B., DIANE, S., SYLVERE, B. and MUSUMECI, S. (2006). Network of sexual contacts and sexually transmitted HIV infection in Burkina Faso. *J. Med. Virol.* **78** 724–729.

LAZER, D., RUBINEAU, B., CHETKOVICH, C., KATZ, N. and NEBLO, M. (2010). The coevolution of networks and political attitudes. *Polit. Commun.* **27** 248–274.

LYONS, R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics*, *Politics*, *and Policy* **2**.

NOEL, H. and NYHAN, B. (2011). The unfriending problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Soc. Netw.* **33** 211–218.
O'BRIEN, K. L. and DAGAN, R. (2003). The potential indirect effect of conjugate pneumococcal vaccines. *Vaccine* **21** 1815–1825.
OGBURN, E. L. and VANDERWEELE, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29** 559–578. MR3300359
OGBURN, E. L., VANDERWEELE, T. J. and CHRISTAKIS, N. A. (2017). Supplement to "Vaccines, contagion, and social networks." DOI:10.1214/17-AOAS1023SUPP.
OSTERHOLM, M. T., KELLEY, N. S., SOMMER, A. and BELONGIA, E. A. (2012). Efficacy and effectiveness of influenza vaccines: A systematic review and meta-analysis. *Lancet, Infect. Dis.* **12** 36–44.
PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* 411–420.
ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.
ROBINS, J. M. and RICHARDSON, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology*: *Finding the Determinants of Disorders and Their Cures* (P. Shrout, ed.). Oxford Univ. Press.
ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537
ROSENQUIST, J. N., MURABITO, J., FOWLER, J. H. and CHRISTAKIS, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Ann. Intern. Med.* **152** 426–433.
SHALIZI, C. R. (2012). Comment on "Why and when 'flawed' social network analyses still yield valid tests of no contagion." *Statistics, Politics, and Policy* **3**.
SHALIZI, C. R. and THOMAS, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* **40** 211–239. MR2767833
TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. In *Proceedings of the* 30*th International Conference on Machine Learning* 1489–1497.
UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 329–337. ACM.
VALERI, L. and VANDERWEELE, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* **18** 137–150.
VANDERWEELE, T. J. (2011). Sensitivity analysis for contagion effects in social networks. *Sociol. Methods Res.* **40** 240–255. MR2767834
VANDERWEELE, T. J., OGBURN, E. L. and TCHETGEN TCHETGEN, E. J. (2012). Why and when "flawed" social network analyses still yield valid tests of no contagion. *Statistics, Politics, and Policy* **3** 1–11.
VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011a). Bounding the infectiousness effect in vaccine trials. *Epidemiology* **22** 686–693.
VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011b). Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Probab. Lett.* **81** 861–869. MR2793754
VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J. and HALLORAN, M. E. (2012). Components of the indirect effect in vaccine trials: Identification of contagion and infectiousness effects. *Epidemiology* **23** 751–761.
VAN DER LAAN, M. J. (2012). Causal inference for networks. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 300.
YANG, Y., SUGIMOTO, J. D., HALLORAN, M. E., BASTA, N. E., CHAO, D. L., MATRAJT, L., POTTER, G., KENAH, E. and LONGINI, I. M. (2009). The transmissibility and control of pandemic influenza A (H1N1) virus. *Science* **326** 729–733.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 949–966
DOI: 10.1214/17-AOAS1024

SUBGROUP INFERENCE FOR MULTIPLE TREATMENTS AND MULTIPLE ENDPOINTS IN AN ALZHEIMER'S DISEASE TREATMENT TRIAL

BY PATRICK SCHNELL*, QI TANG†,‡, PETER MÜLLER§ AND BRADLEY P. CARLIN*

University of Minnesota, AbbVie, Inc.†, Sanofi, Inc.‡ and University of Texas at Austin§*

Many new experimental treatments outperform the current standard only for a subset of the population. Subgroup identification methods provide estimates for the population subset which benefits most from treatment. However, when more than two treatments and multiple endpoints are under consideration, there are many possible requirements for a particular treatment to be beneficial. In this paper, we adapt notions of decision-theoretic admissibility to the context of evaluating treatments in such trials. As an explicit demonstration of admissibility concepts, we combine our approach with the method of credible subgroups, which in the case of a single outcome and treatment comparison provides Bayesian bounds on the benefiting subpopulation. We investigate our methods' performance via simulation, and apply them to a recent dataset from an Alzheimer's disease treatment trial. Our results account for multiplicity while showing patient covariate profiles that are (or are not) likely to be associated with treatment benefit, and are thus useful in their own right or as a guide to patient enrollment in a second stage study.

REFERENCES

ALMIRALL, D., LIZOTTE, D. J. and MURPHY, S. A. (2012). Comment: "Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer". *J. Amer. Statist. Assoc.* **107** 509–512. MR2980061

BERGER, J. O., WANG, X. and SHEN, L. (2014). A Bayesian approach to subgroup identification. *J. Biopharm. Statist.* **24** 110–129. MR3196130

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

BURNS, A. and ILIFFE, S. (2009). Alzheimer's disease. *BMJ* **338** b158.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

DIXON, D. O. and SIMON, R. (1991). Bayesian subset analysis. *Biometrics* 871–881.

FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689

Key words and phrases. Bayesian inference, clinical trials, heterogeneous treatment effect, linear model, simultaneous inference, subgroup identification.

FREIDLIN, B., JIANG, W. and SIMON, R. (2010). The cross-validated adaptive signature design. *Clin. Cancer Res.* **16** 691–698.

FREIDLIN, B. and SIMON, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin. Cancer Res.* **11** 7872–7878.

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

LABER, E. B., LIZOTTE, D. J. and FERGUSON, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70** 53–61. MR3251666

LIZOTTE, D. J., BOWLING, M. and MURPHY, S. A. (2012). Linear fitted-Q iteration with multiple reward functions. *J. Mach. Learn. Res.* **13** 3253–3295. MR3005887

LIZOTTE, D. J. and LABER, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. *J. Mach. Learn. Res.* **17** 1–28. MR3595145

NIMBLE DEVELOPMENT TEAM (2015). NIMBLE: An R Package for Programming with BUGS models, Version 0.4.

PEACE, K. E. and CHEN, D.-G. D. (2010). *Clinical Trial Methodology*. Chapman & Hall/CRC, Boca Raton, FL.

POCOCK, S. J., ASSMANN, S. E., ENOS, L. E. and KASTEN, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Stat. Med.* **21** 2917–2930.

SCHNELL, P. M., TANG, Q., OFFEN, W. W. and CARLIN, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72** 1026–1036.

SCHNELL, P. M., TANG, Q., MÜLLER, P. and CARLIN, B. P. (2017). Supplement to "Subgroup inference for multiple treatments and multiple endpoints in an Alzheimer's disease treatment trial." DOI:10.1214/17-AOAS1024SUPP.

SIVAGANESAN, S., LAUD, P. W. and MÜLLER, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya urn scheme. *Stat. Med.* **30** 312–323. MR2758864

THALL, P. F., SUNG, H.-G. and ESTEY, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Amer. Statist. Assoc.* **97** 29–39. MR1947271

THALL, P. F., LOGOTHETIS, C., PAGLIARO, L. C., WEN, S., BROWN, M. A., WILLIAMS, D. and MILLIKAN, R. E. (2007). Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. *J. Natl. Cancer Inst.* **99** 1613–1622.

UUSIPAIKKA, E. (1983). Exact confidence bands for linear regression over intervals. *J. Amer. Statist. Assoc.* **78** 638–644. MR0721213

XU, Y., TRIPPA, L., MÜLLER, P. and JI, Y. (2016). Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Stat. Biosci.* **8** 159–180.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 967–991
DOI: 10.1214/17-AOAS1026

QUANTIFICATION OF MULTIPLE TUMOR CLONES USING GENE ARRAY AND SEQUENCING DATA

BY YICHEN CHENG[*,1], JAMES Y. DAI[1], THOMAS G. PAULSON[2], XIAOYU WANG, XIAOHONG LI[2], BRIAN J. REID[2] AND CHARLES KOOPERBERG[1]

Fred Hutchinson Cancer Research Center

Cancer development is driven by genomic alterations, including copy number aberrations. The detection of copy number aberrations in tumor cells is often complicated by possible contamination of normal stromal cells in tumor samples and intratumor heterogeneity, namely the presence of multiple clones of tumor cells. In order to correctly quantify copy number aberrations, it is critical to successfully de-convolute the complex structure of the genetic information from tumor samples. In this article, we propose a general Bayesian method for estimating copy number aberrations when there are normal cells and potentially more than one tumor clones. Our method provides posterior probabilities for the proportions of tumor clones and normal cells. We incorporate prior information on the distribution of the copy numbers to prioritize biologically more plausible solutions and alleviate possible identifiability issues that have been observed by many researchers. Our model is flexible and can work for both SNP array and next-generation sequencing data. We compare our method to existing ones and illustrate the advantage of our approach in multiple datasets.

REFERENCES

ANDOR, N., HARNESS, J. V., MÜLLER, S., MEWES, H. W. and PETRITSCH, C. (2014). EXPANDS: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30** 50–60.

ATTIYEH, E. F., DISKIN, S. J., ATTIYEH, M. C., MOSSÉ, Y. P., HOU, C., JACKSON, E. M., KIM, C., GLESSNER, J., HAKONARSON, H., BIEGEL, J. A. and MARIS, J. M. (2009). Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res.* **19** 276–283.

BAO, L., PU, M. and MESSER, K. (2014). AbsCN-seq: A statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* **30** 1056–1063.

BEROUKHIM, R., MERMEL, C. H., PORTER, D., WEI, G., RAYCHAUDHURI, S., DONOVAN, J., BARRETINA, J., BOEHM, J. S., DOBSON, J., URASHIMA, M., MC HENRY, K. T., PINCHBACK, R. M., LIGON, A. H., CHO, Y. J., HAERY, L., GREULICH, H., REICH, M., WINCKLER, W., LAWRENCE, M. S., WEIR, B. A., TANAKA, K. E., CHIANG, D. Y., BASS, A. J., LOO, A. L., HOFFMAN, C., PRENSNER, J., LIEFELD, T., GAO, Q., YECIES, D., SIGNORETTI, S., MAHER, E., KAYE, F. J., SASAKI, H., TEPPER, J. E., FLETCHER, J. A., TABERNERO, J., BASELGA, J., TSAO, M. S., DEMICHELIS, F., RUBIN, M. A., JANNE, P. A.,

Key words and phrases. Copy number aberration, intratumor heterogeneity, identifiability, BIC.

DALY, M. J., NUCERA, C., LEVINE, R. L., EBERT, B. L., GABRIEL, S., RUSTGI, A. K., ANTONESCU, C. R., LADANYI, M., LETAI, A., GARRAWAY, L. A., LODA, M., BEER, D. G., TRUE, L. D., OKAMOTO, A., POMEROY, S. L., SINGER, S., GOLUB, T. R., LANDER, E. S., GETZ, G., SELLERS, W. R. and MEYERSON, M. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* **463** 899–905.

CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W., WEIR, B. A., BEROUKHIM, R., PELLMAN, D., LEVINE, D. A., LANDER, E. S., MEYERSON, M. and GETZ, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.

DE BRUIN, E. C., MCGRANAHAN, N., MITTER, R., SALM, M., WEDGE, D. C., YATES, L., JAMAL-HANJANI, M., SHAFI, S., MURUGAESU, N., ROWAN, A. J., GRÄNROOS, E., MUHAMMAD, M. A., HORSWELL, S., GERLINGER, M., VARELA, I., JONES, D., MARSHALL, J., VOET, T., LOO, P. V., RASSL, D. M., RINTOUL, R. C., JANES, S. M., LEE, S. M., FORSTER, M., AHMAD, T., LAWRENCE, D., FALZON, M., CAPITANIO, A., HARKINS, T. T., LEE, C. C., TOM, W., TEEFE, E., CHEN, S.-C., BEGUM, S., RABINOWITZ, A., PHILLIMORE, B., SPENCER-DENE, B., STAMP, G., SZALLASI, Z., MATTHEWS, N., STEWART, A., CAMPBELL, P. and SWANTON, C. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346** 251–256.

GAZDAR, A. F., KURVARI, V., VIRMANI, A., GOLLAHON, L., SAKAGUCHI, M., WESTERFIELD, M., KODAGODA, D., STASNY, V., CUNNINGHAM, H. T., WISTUBA, I. I., TOMLINSON, G., TONK, V., ASHFAQ, R., LEITCH, A. M., MINNA, J. D. and SHAY, J. W. (1998). Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int. J. Cancer* **78** 766–774.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. and SWANTON, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366** 883–892.

GU, J., AJANI, J. A., HAWK, E. T., YE, Y., LEE, J. H., BHUTANI, M. S., HOFSTETTER, W. L., SWISHER, S. G., WANG, K. K. and WU, X. (2010). Genome-wide catalogue of chromosomal aberrations in barrett's esophagus and esophageal adenocarcinoma: A high-density single nucleotide polymorphism array analysis. *Cancer Prev. Res.* **3** 1176–1186.

LARSON, N. B. and FRIDLEY, B. L. (2013). PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29** 1888–1889.

LI, X., GALIPEAU, P. C., PAULSON, T. G., SANCHEZ, C. A., ARNAUDO, J., LIU, K., SATHER, C. L., KOSTADINOV, R. L., ODZE, R. D., KUHNER, M. K., MALEY, C. C., SELF, S. G., VAUGHAN, T. L., BLOUNT, P. L. and REID, B. J. (2014). Temporal and spatial evolution of somatic chromosomal alterations: A case-cohort study of Barrett's esophagus. *Cancer Prev. Res.* **7** 114–127.

MICHOR, F. and POLYAK, K. (2010). The origins and implications of intratumor heterogeneity. *Cancer Prev. Res.* **3** 1361–1364.

OESPER, L., MAHMOODY, A. and RAPHAEL, B. J. (2013). THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14** R80.

OESPER, L., SATAS, G. and RAPHAEL, B. J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30** 3532–3540.

OLSHEN, A. B., BENGTSSON, H., NEUVIAL, P., SPELLMAN, P. T., OLSHEN, R. A. and SESHAN, V. E. (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics* **27** 2038–2046.

REID, B. J., LI, X., GALIPEAU, P. C. and VAUGHAN, T. L. (2010). Barrett's oesophagus and oesophageal adenocarcinoma: Time for a new synthesis. *Nat. Rev. Cancer* **10** 87–101.

STAAF, J., LINDGREN, D., VALLON-CHRISTERSSON, J., ISAKSSON, A., GÖRANSSON, H., JULIUSSON, G., ROSENQUIST, R., HÖGLUND, M., BORG, Å. and RINGNÉR, M. (2008). Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* **9** R136.

VAN LOO, P., NORDGARD, S. H., LINGJÆRDE, O. C., RUSSNES, H. G., TYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C. M., BØRRESEN-DALE, A. and KRISTENSEN, V. N. (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107** 16910–16915.

VOLM, M., MATTERN, J., SONKA, J., VOGT-SCHADEN, M. and WAYSS, K. (1985). DNA distribution in non-small-cell lung carcinomas and its relationship to clinical behavior. *Cytometry* **6** 348–56.

WANG, K. K., SAMPLINER, R. E. and PRACTICE PARAMETERS COMMITTEE OF THE AMERICAN COLLEGE OF GASTROENTEROLOGY (2008). Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus. *Am. J. Gastroenterol.* **103** 788–797.

XU, Y., MÜLLER, P., YUAN, Y., GULUKOTA, K. and JI, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *J. Amer. Statist. Assoc.* **110** 503–514. MR3367243

YAU, C. (2013). OncoSNP-SEQ: A statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29** 2482–2484.

YAU, C., MOURADOV, D., JORISSEN, R. N., COLELLA, S., MIRZA, G., STEERS, G., HARRIS, A., RAGOUSSIS, J., SIEBER, O. and HOLMES, C. C. (2010). A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol.* **11** R92.

YU, Z., LIU, Y., SHEN, Y., WANG, M. and LI, A. (2014). CLImAT: Accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics* **30** 2576–2583.

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

ZHANG, J., FUJIMOTO, J., ZHANG, J., WEDGE, D. C., SONG, X., ZHANG, J., SETH, S., CHOW, C.-W., CAO, Y., GUMBS, C., GOLD, K. A., KALHOR, N., LITTLE, L., MAHADESHWAR, H., MORAN, C., PROTOPOPOV, A., SUN, H., TANG, J., WU, X., YE, Y., WILLIAM, W. N., LEE, J. J., HEYMACH, J. V., HONG, W. K., SWISHER, S., WISTUBA, I. I. and FUTREAL, P. A. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346** 256–259.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 992–1010
DOI: 10.1214/17-AOAS1030

GENERALIZED MAHALANOBIS DEPTH IN POINT PROCESS AND ITS APPLICATION IN NEURAL CODING

BY SHUYI LIU AND WEI WU

Florida State University

In this paper, we propose to generalize the notion of depth in temporal point process observations. The new depth is defined as a weighted product of two probability terms: (1) the number of events in each process, and (2) the center-outward ranking on the event times conditioned on the number of events. In this study, we adopt the Poisson distribution for the first term and the Mahalanobis depth for the second term. We propose an efficient bootstrapping approach to estimate parameters in the defined depth. In the case of Poisson process, the observed events are order statistics where the parameters can be estimated robustly with respect to sample size. We demonstrate the use of the new depth by ranking realizations from a Poisson process. We also test the new method in classification problems using simulations as well as real neural spike train data. It is found that the new framework provides more accurate and robust classifications as compared to commonly used likelihood methods.

REFERENCES

BERRAR, D. P., DUBITZKY, W. and GRANZOW, M. (2009). *A Practical Approach to Microarray Data Analysis*. Springer, Berlin.

BOX, G. E. P., HUNTER, W. G. and HUNTER, J. S. (1978). *Statistics for Experimenters*: *An Introduction to Design*, *Data Analysis*, *and Model Building*. Wiley, New York. MR0483116

BROWN, E. N., BARBIERI, R., VENTURA, V., KASS, R. E. and FRANK, L. M. (2001). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput*. **14** 325–346.

CASCOS, I. (2007). The expected convex hull trimmed regions of a sample. *Comput*. *Statist*. **22** 557–569. MR2358426

DIEZ, D. M., SCHOENBERG, F. P. and WOODY, C. D. (2012). Algorithms for computing spike time distance and point process prototypes with application to feline neuronal responses to acoustic stimuli. *J*. *Neurosci*. *Methods* **203** 186–192.

DRAZEK, L. C. (2013). Intensity estimation for Poisson Processes. The University of Leeds, School of Mathematics.

DYCKERHOFF, R., MOSLER, K. and KOSHEVOY, G. (1996). Zonoid data depth: Theory and computation. In *OMPSTAT*: *Proceedings in Computational Statistics* 12*th Symposium held in Barcelona*, *Spain*, 1996 235–240. Physica-Verlag, Heidelberg.

DYCKERHOFF, R. and MOSLER, K. (2011). Weighted-mean trimming of multivariate data. *J*. *Multivariate Anal*. **102** 405–421. MR2755006

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann*. *Statist*. **7** 1–26. MR0515681

Key words and phrases. Point process, Mahalanobis depth, Poisson process, neural coding, spike train.

FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London. MR1699953

JULIENNE, H. and HOUGHTON, C. (2013). A simple algorithm for averaging spike trains. *J. Math. Neurosci.* **3** Art. 3, 14. MR3042443

KARLIN, S. (1966). *A First Course in Stochastic Processes*. Academic Press, New York. MR0208657

KASS, R. E., VENTURA, V. and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J. Neurophysiol.* **94** 8–25.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford Univ. Press, New York. MR1419991

LAWHERN, V., NIKONOV, A. A., WU, W. and CONTRARES, R. J. (2011). Spike rate and spike timing contributions to coding taste quality information in rat periphery. *Frontiers in Integrative Neuroscience* **5** 1–14.

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400

LIU, R. Y. and SINGH, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.* **88** 252–260. MR1212489

LÓPEZ-PINTADO, S. and ROMO, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.* **104** 718–734. MR2541590

MATEU, J., SCHOENBERG, F. P., DIEZ, D. M., GONZÁLEZ, J. A. and LU, W. (2015). On measures of dissimilarity between point patterns: Classification based on prototypes and multidimensional scaling. *Biom. J.* **57** 340–358. MR3324443

MOGHADAM, S. A. and PAZIRA, H. (2011). The relations among the order statistics of uniform distribution. *Trends in Applied Science Research* **6** 719–723.

MOSLER, K. and POLYAKOVA, Y. (2012). General notions of depth for functional data. Available at arXiv:1208.1981.

ROBERT, C. (1995). Simulation of truncated normal variables. *Stat. Comput.* 121–125.

ROSS, S. M. (1983). *Stochastic Processes*. Wiley, New York. MR0683455

TUKEY, J. W. (1975). Mathematics and the picturing of data. *International Congress of Mathematicians* **2** 523–531. MR0426989

VICTOR, J. D. and PURPURA, K. P. (1997). Metric-space analysis of spike trains: Theory, algorithms and application. *Network* **8** 127–164.

WESOLOWSKI, S., CONTRERAS, R. J. and WU, W. (2015). A new framework for Euclidean summary statistics in the neural spike train space. *Ann. Appl. Stat.* **9** 1278–1297. MR3418723

WU, W. and SRIVASTAVA, A. (2011). An information-geometric framework for statistical inferences in the neural spike train space. *J. Comput. Neurosci.* **31** 725–748. MR2864743

WU, W. and SRIVASTAVA, A. (2013). Estimating summary statistics in the spike-train space. *J. Comput. Neurosci.* **34** 391–410. MR3061973

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1011–1039
DOI: 10.1214/17-AOAS1033

INTEGRATIVE SPARSE *K*-MEANS WITH OVERLAPPING GROUP LASSO IN GENOMIC APPLICATIONS FOR DISEASE SUBTYPE DISCOVERY

BY ZHIGUANG HUO[1] AND GEORGE TSENG[1]

University of Pittsburgh

Cancer subtypes discovery is the first step to deliver personalized medicine to cancer patients. With the accumulation of massive multi-level omics datasets and established biological knowledge databases, omics data integration with incorporation of rich existing biological knowledge is essential for deciphering a biological mechanism behind the complex diseases. In this manuscript, we propose an integrative sparse *K*-means (IS-*K*means) approach to discover disease subtypes with the guidance of prior biological knowledge via sparse overlapping group lasso. An algorithm using an alternating direction method of multiplier (ADMM) will be applied for fast optimization. Simulation and three real applications in breast cancer and leukemia will be used to compare IS-*K*means with existing methods and demonstrate its superior clustering accuracy, feature selection, functional annotation of detected molecular features and computing efficiency.

REFERENCES

ABRAMSON, V. G., LEHMANN, B. D., BALLINGER, T. J. and PIETENPOL, J. A. (2015). Subtyping of triple-negative breast cancer: Implications for therapy. *Cancer* **121** 8–16.

BALGOBIND, B. V., VAN DEN HEUVEL-EIBRINK, M. M., DE MENEZES, R. X., REINHARDT, D., HOLLINK, I. H. I. M., ARENTSEN-PETERS, S. T. J. C. M., VAN WERING, E. R., KASPERS, G. J. L., CLOOS, J., DE BONT, E. S. J. M., CAYUELA, J. M., BARUCHEL, A., MEYER, C., MARSCHALEK, R., TRKA, J., STARY, J., BEVERLOO, H. B., PIETERS, R., ZWAAN, C. M. and DEN BOER, M. L. (2010). Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica* **96** 221–230.

BASS, A., THORSSON, V., SHMULEVICH, I., REYNOLDS, S., MILLER, M., BERNARD, B., HINOUE, T., LAIRD, P., CURTIS, C., SHEN, H. et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513** 202–209.

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.

CURTIS, C., SHAH, S. P., CHIN, S.-F., TURASHVILI, G., RUEDA, O. M., DUNNING, M. J., SPEED, D., LYNCH, A. G., SAMARAJIWA, S., YUAN, Y. et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486** 346–352.

DOMANY, E. (2014). Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. *Cancer Res.* **74** 4612–4621.

Key words and phrases. Cancer subtype, omics integrative analysis, overlapping group lasso, ADMM.

DUDOIT, S. and FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol*. **3** 1–21.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** 14863–14868.

FAN, X. and KURGAN, L. (2015). Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief. Bioinform*. **16** 780–794.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

HE, B. S., YANG, H. and WANG, S. L. (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *J. Optim. Theory Appl*. **106** 337–356. MR1788928

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist*. **38** 2282–2313. MR2676890

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

HUO, Z. and TSENG, G. (2017). Supplement to "Integrative sparse K-means with overlapping group lasso in genomic applications for disease subtype discovery." DOI:10.1214/17-AOAS1033SUPP.

HUO, Z., DING, Y., LIU, S., OESTERREICH, S. and TSENG, G. (2016). Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *J. Amer. Statist. Assoc*. **111** 27–42. MR3494636

JACCARD, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaud. Sci. Nat*. **37** 547–579.

JACOB, L., OBOZINSKI, G. and VERT, J. P. (2009). Group lasso with overlap and graph lasso. In *ICML* '09: *Proceedings of the* 26*th Annual International Conference on Machine Learning* 433–440. ACM, New York.

KAUFMAN, L. and ROUSSEEUW, P. (1987). *Clustering by Means of Medoids*. North-Holland, Amsterdam.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data*: *An Introduction to Cluster Analysis*. Wiley, New York. MR1044997

KIM, E.-Y., KIM, S.-Y., ASHLOCK, D. and NAM, D. (2009). MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinform*. **10** 260.

KOBOLDT, D. C., FULTON, R. S., MCLELLAN, M. D., SCHMIDT, H., KALICKI-VEIZER, J., MCMICHAEL, J. F., FULTON, L. L., DOOLING, D. J., DING, L., MARDIS, E. R. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

KOHLMANN, A., KIPPS, T. J., RASSENTI, L. Z., DOWNING, J. R., SHURTLEFF, S. A., MILLS, K. I., GILKES, A. F., HOFMANN, W.-K., BASSO, G., DELL'ORTO, M. C., FOÀ, R., CHIARETTI, S., VOS, J. D., RAUHUT, S., PAPENHAUSEN, P. R., HERNÁNDEZ, J. M., LUMBRERAS, E., YEOH, A. E., KOAY, E. S., LI, R., LIU, W., WILLIAMS, P. M., WIECZOREK, L. and HAFERLACH, T. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: The microarray innovations in LEukemia study prephase. *Br. J. Haematol*. **142** 802–807.

LEHMANN, B. D., BAUER, J. A., CHEN, X., SANDERS, M. E., CHAKRAVARTHY, A. B., SHYR, Y. and PIETENPOL, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest*. **121** 2750.

LOCK, E. F. and DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29** 2610–2616.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif*., 1965/66) 281–297. Univ. California Press, Berkeley, CA. MR0214227

MAITRA, R. and RAMLER, I. P. (2009). Clustering in the presence of scatter. *Biometrics* **65** 341–352. MR2751457

MCLACHLAN, G. J., BEAN, R. and PEEL, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18** 413–422.

MILLIGAN, G. W. and COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50** 159–179.

PARKER, J. S., MULLINS, M., CHEANG, M. C., LEUNG, S., VODUC, D., VICKERY, T., DAVIES, S., FAURON, C., HE, X., HU, Z. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27** 1160–1167.

PARSONS, D. W., JONES, S., ZHANG, X., LIN, J. C.-H., LEARY, R. J., ANGENENDT, P., MANKOO, P., CARTER, H., SIU, I.-M., GALLIA, G. L. et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* **321** 1807–1812.

QIN, Z. S. (2006). Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics* **22** 1988–1997.

RAMASAMY, A., MONDRY, A., HOLMES, C. C. and ALTMAN, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5** 1320–1333.

RICHARDSON, S., TSENG, G. C. and SUN, W. (2016). Statistical methods in integrative genomics. *Annu. Rev. Statist. Appl.* **3** 181–209.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MULLER-HERMELINK, H. K., SMELAND, E. B., GILTNANE, J. M. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947.

SADANANDAM, A., LYSSIOTIS, C. A., HOMICSKO, K., COLLISSON, E. A., GIBB, W. J., WULLSCHLEGER, S., OSTOS, L. C. G., LANNON, W. A., GROTZINGER, C., DEL RIO, M. et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19** 619–625.

SHEN, R., OLSHEN, A. B. and LADANYI, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.

SHEN, K. and TSENG, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* **26** 1316–1323.

SIMON, R. (2005). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J. Natl. Cancer Inst.* **97** 866–867.

SIMON, R., RADMACHER, M. D., DOBBIN, K. and MCSHANE, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95** 14–18.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

SWIFT, S., TUCKER, A., VINCIOTTI, V., MARTIN, N., ORENGO, C., LIU, X. and KELLAM, P. (2004). Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.* **5** R94.

TIBSHIRANI, R. and WALTHER, G. (2005). Cluster validation by prediction strength. *J. Comput. Graph. Statist.* **14** 511–528. MR2170199

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503

TOTHILL, R. W., TINKER, A. V., GEORGE, J., BROWN, R., FOX, S. B., LADE, S., JOHNSON, D. S., TRIVETT, M. K., ETEMADMOGHADAM, D., LOCANDRO, B. et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14** 5198–5208.

TSENG, G. C. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* **23** 2247–2255.

TSENG, G., GHOSH, D. and FEINGOLD, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*. **40** 3785–3799.

TSENG, G. C. and WONG, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61** 10–16. MR2129196

VERHAAK, R. G., WOUTERS, B. J., ERPELINCK, C. A., ABBAS, S., BEVERLOO, H. B., LUGTHART, S., LÖWENBERG, B., DELWEL, R. and VALK, P. J. (2009). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94** 131–134.

VERHAAK, R. G., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T., MESIROV, J. P. 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, S. L. and LIAO, L. Z. (2001). Decomposition method with a variable parameter for a class of monotone variational inequality problems. *J. Optim. Theory Appl*. **109** 415–429. MR1834183

WITKOS, T. M., KOSCIANSKA, E. and KRZYZOSIAK, W. J. (2011). Practical aspects of microRNA target prediction. *Curr. Mol. Med*. **11** 93–109.

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc*. **105** 713–726. MR2724855

XIE, B., PAN, W. and SHEN, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electron. J. Stat*. **2** 168–212. MR2386092

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

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc*. **101** 1418–1429. MR2279469

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1040–1062
DOI: 10.1214/17-AOAS1035

MULTILEVEL MODELS WITH STOCHASTIC VOLATILITY FOR REPEATED CROSS-SECTIONS: AN APPLICATION TO TRIBAL ART PRICES

BY SILVIA CAGNONE[1], SIMONE GIANNERINI AND LUCIA MODUGNO[1]

University of Bologna

In this paper, we introduce a multilevel specification with stochastic volatility for repeated cross-sectional data. Modelling the time dynamics in repeated cross sections requires a suitable adaptation of the multilevel framework where the individuals/items are modelled at the first level whereas the time component appears at the second level. We perform maximum likelihood estimation by means of a nonlinear state space approach combined with Gauss–Legendre quadrature methods to approximate the likelihood function. We apply the model to the first database of tribal art items sold in the most important auction houses worldwide. The model allows to account properly for the heteroscedastic and autocorrelated volatility observed and has superior forecasting performance. Also, it provides valuable information on market trends and on predictability of prices that can be used by art markets stakeholders.

REFERENCES

AGNELLO, R. and PIERCE, R. (1996). Financial returns, price determinants, and genre effects in American art investment. *J. Cult. Econ.* **20** 359–383.

BALLESTEROS, T. (2011). Efficiency tests in the art market using cointegration and the error correction model. *Social Science Research Network*. DOI:http://dx.doi.org/10.2139/ssrn.1696785.

BARTOLUCCI, F. and DE LUCA, G. (2001). Maximum likelihood estimation of a latent variable time-series model. *Appl. Stoch. Models Bus. Ind.* **17** 5–17. MR1819006

BAUMOL, W. (1986). Unnatural value: Or art investment as floating crap game. *Am. Econ. Rev.* **76** 10–14.

BIORDI, M. and CANDELA, G. (2007). L'arte etnica: Tra cultura e mercato. *Skira*.

BOCART, F. Y. R. P. and HAFNER, C. M. (2012). Econometric analysis of volatile art markets. *Comput. Statist. Data Anal.* **56** 3091–3104. MR2943883

BOCART, F. Y. R. P. and HAFNER, C. M. (2015). Volatility of price indices for heterogeneous goods with applications to the fine art market. *J. Appl. Econometrics* **30** 291–312. MR3322720

BOUND, J., JAEGER, D. and BAKER, R. (1995). Problems with instrumental variables estimation when the correlation between instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.

BOX-STEFFENSMEIER, J. M., DE BOEF, S. and LIN, T.-M. (2004). The dynamics of the partisan gender gap. *Am. Polit. Sci. Rev.* **98** 515–528.

BROWNE, W. and GOLDSTEIN, H. (2010). MCMC sampling for a multilevel model with nonindependent residuals within and between cluster units. *J. Educ. Behav. Stat.* **35** 453–473.

Key words and phrases. Multilevel model, hedonic regression model, dependent random effects, stochastic volatility, autoregression.

CAGNONE, S. and BARTOLUCCI, F. (2017). Adaptive quadrature for maximum likelihood estimation of a class of dynamic latent variable models. *Comput. Econ.* **49** 599–622.

CAGNONE, S., GIANNERINI, S. and MODUGNO, L. (2017). Supplement to "Multilevel models with stochastic volatility for repeated cross-sections: An application to tribal art prices." DOI:10.1214/17-AOAS1035SUPP.

CANDELA, G., CASTELLANI, M. and PATTITONI, P. (2012). Tribal art market: Signs and signals. *J. Cult. Econ.* **36** 289–308.

CANDELA, G., CASTELLANI, M. and PATTITONI, P. (2013). Reconsidering psychic return in art investments. *Econom. Lett.* **118** 351–354.

CHANEL, O. (1995). Is art market behaviour predictable? *Eur. Econ. Rev.* **39** 519–527.

CHANEL, O., GÉRARD-VARET, L. and GINSBURGH, V. (1996). The relevance of hedonic price indices. The case of paintings. *J. Cult. Econ.* **20** 1–24.

COLLINS, A., SCORCU, A. and ZANOLA, R. (2009). Reconsidering hedonic art price indexes. *Econom. Lett.* **104** 57–60. MR2542857

DEATON, A. (1985). Panel data from time series of cross sections. *J. Econometrics* **30** 109–126.

DIPRETE, T. and GRUSKY, D. (1990). The multilevel analysis of trends with repeated cross-sectional data. *Sociol. Method.* **20** 337–368.

DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. MR3014996

FRIDMAN, M. and HARRIS, L. (1998). A maximum likelihood approach for non-Gaussian stochastic volatility models. *J. Bus. Econom. Statist.* **16** 284–291.

GIANNERINI, S. (2015). tseriesEntropy: Entropy Based Analysis and Tests for Time Series. R package version 0.5-13.

GIANNERINI, S., MAASOUMI, E. and BEE DAGUM, E. (2015). Entropy testing for nonlinear serial dependence in time series. *Biometrika* **102** 661–675. MR3394282

GINSBURGH, V. and JEANFILS, P. (1995). Long-term comovements in international markets for paintings. *Eur. Econ. Rev.* **39** 538–548.

GOETZMANN, W. (1993). Accounting for taste: Art and financial markets over three centuries. *Am. Econ. Rev.* **83** 1370–1376.

GOETZMANN, W. (1995). The informational efficiency of the art market. *Manage. Finance* **21** 25–34.

GOETZMANN, W., MAMONOVA, E. and SPAENJERS, C. (2014). The economics of aesthetics and three centuries of art price records. Working Paper 20440, National Bureau of Economic Research.

GOLDSTEIN, H. (2010). *Multilevel Statistical Models*, 4th ed. Wiley, Chichester.

GRANGER, C. W., MAASOUMI, E. and RACINE, J. (2004). A dependence metric for possibly nonlinear processes. *J. Time Series Anal.* **25** 649–669. MR2086354

HODGSON, D. and VORKINK, K. (2004). Asset pricing theory and the valuation of Canadian paintings. *Canadian Journal of Economics/Revue canadienne déconomique* **37** 629–655.

JOANES, D. and GILL, C. (1998). Comparing measures of sample skewness and kurtosis. *J. R. Stat. Soc., Ser. D Stat.* **47** 183–189.

JONES, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach. Monographs on Statistics and Applied Probability* **47**. Chapman & Hall, London. MR1293123

KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.* **82** 1032–1063. MR0922169

LEBO, M. and WEBER, C. (2015). An effective approach to the repeated cross-sectional design. *Amer. J. Polit. Sci.* **59** 242–258.

LEVENE, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, ed.) 278–292. Stanford Univ. Press, Stanford, CA. MR0120709

LOCATELLI BIEY, M. and ZANOLA, R. (2005). The market for picasso prints: A hybrid model approach. *J. Cult. Econ.* **29** 127–136.

MACKUEN, M., ERIKSON, R. and STIMSON, J. (1992). Peasants or bankers? The American electorate and the U.S. economy. *Am. Polit. Sci. Rev.* **86** 597–611.

MODUGNO, L., CAGNONE, S. and GIANNERINI, S. (2015). A multilevel model with autoregressive components for the analysis of tribal art prices. *J. Appl. Stat.* **42** 2141–2158. MR3373724

MODUGNO, L. and GIANNERINI, S. (2015). The wild bootstrap for multilevel models. *Comm. Statist. Theory Methods* **44** 4812–4825. MR3424810

MOFFITT, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *J. Econometrics* **59** 99–123.

ROSEN, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *J. Polit. Econ.* **82** 34–55.

SCOTT, A. J. and SMITH, T. M. F. (1974). Analysis of repeated surveys using time series methods. *J. Amer. Statist. Assoc.* **69** 674–678.

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL. MR2059021

TANIZAKI, H. and MARIANO, R. (1998). Nonlinear and nonnormal state-space modeling with Monte-Carlo stochastic simulations. *J. Econometrics* **83** 263–290.

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

XU, R. (2003). Measuring explained variation in linear mixed effects models. *Stat. Med.* **22** 3527–3541.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1063–1084
DOI: 10.1214/17-AOAS1036

FLEXIBLE RISK PREDICTION MODELS FOR LEFT OR INTERVAL-CENSORED DATA FROM ELECTRONIC HEALTH RECORDS

BY NOORIE HYUN*, LI C. CHEUNG*, QING PAN†, MARK SCHIFFMAN* AND HORMUZD A. KATKI*

*National Cancer Institute** *and George Washington University*†

Electronic health records are a large and cost-effective data source for developing risk-prediction models. However, for screen-detected diseases, standard risk models (such as Kaplan–Meier or Cox models) do not account for key issues encountered with electronic health record data: left-censoring of pre-existing (prevalent) disease, interval-censoring of incident disease, and ambiguity of whether disease is prevalent or incident when definitive disease ascertainment is not conducted at baseline. Furthermore, researchers might conduct novel screening tests only on a complex two-phase subsample. We propose a family of weighted mixture models that account for left/interval-censoring and complex sampling via inverse-probability weighting in order to estimate current and future absolute risk: we propose a weakly-parametric model for general use and a semiparametric model for checking goodness of fit of the weakly-parametric model. We demonstrate asymptotic properties analytically and by simulation. We used electronic health records to assemble a cohort of 33,295 human papillomavirus (HPV) positive women undergoing cervical cancer screening at Kaiser Permanente Northern California (KPNC) that underlie current screening guidelines. The next guidelines would focus on HPV typing tests, but reporting 14 HPV types is too complex for clinical use. National Cancer Institute along with KPNC conducted a HPV typing test on a complex subsample of 9258 women in the cohort. We used our model to estimate the risk due to each type and grouped the 14 types (the 3-year risk ranges 21.9–1.5) into 4 risk-bands to simplify reporting to clinicians and guidelines. These risk-bands could be adopted by future HPV typing tests and future screening guidelines.

REFERENCES

BRESLOW, N. E. and WELLNER, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Stat.* **34** 86–102. MR2325244

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009). Improved Horvitz-Thompson estimation of model parameter from two-phase stratified samples: Applications in epidemiology. *Stat. Biosci.* **1** 32–49.

CAI, T. and ZHENG, Y. (2013). Resampling procedures for making inference under nested case-control studies. *J. Amer. Statist. Assoc.* **108** 1532–1544. MR3174727

Key words and phrases. Mixture model, interval censoring, two-phase sampling, B-splines, weighted likelihood, HIV.

CASTLE, P. E., FETTERMAN, B., SCT (ASCP), POITRAS, N., LOREY, T., SHABER, R. and KINNEY, W. (2009). Five-year experience of human papillomavirus DNA and Papanicolaou test cotesting. *Obstetrics & Gynecology* **113** 595–600.

CASTLE, P. E., STOLER, M. H., WRIGHT, JR., T. C., SHARMA, A., WRIGHT, T. L. and BEHRENS, C. M. (2011). Performance of carcinogenic human papillomavirus (HPV) testing and HPV16 or HPV18 genotyping for cervical cancer screening of women aged 25 years and older: A subanalysis of the ATHENA study. *Lancet Oncol*. **12** 880–890.

CHATURVEDI, A. K., KATKI, H. A., HILDESHEIM, A., RODRÍGUEZ, A. C., QUINT, W., SCHIFFMAN, M., VAN DOORN, L. J., PORRAS, C., WACHOLDER, S., GONZALEZ, P. and SHERMAN, M. E. (2011). Human papillomavirus infection with multiple types: Pattern of coinfection and risk of cervical disease. *J. Infect. Dis*. **203** 910–920.

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **34** 187–220. MR0341758

DOREY, F. J., LITTLE, R. J. A. and SCHENKER, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Stat. Med*. **12** 1589–1603.

GRAUBARD, B. I. and KORN, E. L. (1996). Survey inference for subpopulations. *Am. J. Epidemiol*. **144** 102–106.

GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. MR1180321

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc*. **47** 663–685. MR0053460

HUANG, J. and ROSSINI, A. J. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *J. Amer. Statist. Assoc*. **92** 960–967. MR1482126

HUANG, J. and WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics* (D. Y. Lin and T. R. Fleming, eds.) 123–169. Springer, New York.

HYUN, N., CHEUNG, L. C, PAN, Q., SCHIFFMAN, M. and KATKI, H. A (2017). Supplement to "Flexible risk prediction models for left or interval-censored data from electronic health records." DOI:10.1214/17-AOAS1036SUPP.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc*. **53** 457–481. MR0093867

KATKI, H. A., KINNEY, W. K., FETTERMAN, B., LOREY, T., POITRAS, N. E., CHEUNG, L., DEMUTH, F., SCHIFFMAN, M., WACHOLDER, S. and CASTLE, P. E. (2011). Cervical cancer risk for 330,000 women undergoing concurrent HPV testing and cervical cytology in routine clinical practice at a large managed care organization. *Lancet Oncol*. **12** 663–672.

KATKI, H. A., SCHIFFMAN, M., CASTLE, P. E., FETTERMAN, B., POITRAS, N. E., LOREY, T., CHEUNG, L. C., RAINE-BENNETT, T. R., GAGE, J. C. and KINNEY, W. K. (2013). Benchmarking CIN3+ risk as the basis for incorporating HPV and Pap cotesting into cervical screening and management guidelines. *J. Low. Genit. Tract Dis*. **17** S28–S35.

KOVALCHIK, S. A. and PFEIFFER, R. M. (2014). Population-based absolute risk estimation with survey data. *Lifetime Data Anal*. **20** 252–275. MR3181014

LI, C.-S., TAYLOR, J. M. G. and SY, J. P. (2001). Identifiability of cure models. *Statist. Probab. Lett*. **54** 389–395. MR1861384

LUMLEY, T. (2016). Analyses of complex survey samples. Available at https://cran.r-project.org/web/packages/survey/survey.pdf.

MA, S. (2010). Mixed case interval censored data with a cured subgroup. *Statist. Sinica* **20** 1165–1181. MR2730178

MASSAD, L. S., EINSTEIN, M. H., HUH, W. K., KATKI, H. A., KINNEY, W. K., SCHIFFMAN, M., SOLOMON, D., WENTZENSEN, N. and LAWSON, H. W. (2013). 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. *J. Low. Genit. Tract Dis*. **17** S1–S27.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

ODELL, P. M., ANDERSON, K. M. and D'AGOSTINO, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 951–959.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. MR0961262

RÜCKER, R. and MESSERER, D. (1988). Remission duration: An example of interval-censored observations. *Stat. Med.* **7** 1139–1145.

SAEGUSA, T. (2015). Variance estimation under two-phase sampling. *Scand. J. Stat.* **42** 1078–1091. MR3426311

SCHIFFMAN, M., WENTZENSEN, N., WACHOLDER, S., WALTER, K., GAGE, J. C. and CASTLE, P. E. (2011). Human papillomavirus testing in the prevention of cervical cancer. *J. Natl. Cancer Inst.* **103** 368–383.

SCHIFFMAN, M., VAUGHAN, L. M., RAINE-BENNETT, T. R., CASTLE, P. E., KATKI, H. A., GAGE, J. C., FETTERMAN, B., BEFANO, B. and WENTZENSEN, N. (2015). A study of HPV typing for the management of HPV-positive ASC-US cervical cytologic results. *Gynecol. Oncol.* **138** 573–578.

SEN, B. and BANERJEE, M. (2007). A pseudolikelihood method for analyzing interval censored data. *Biometrika* **94** 71–86. MR2307901

SHAO, F., LI, J., MA, S. and LEE, M.-L. T. (2014). Semiparametric varying-coefficient model for interval censored data with a cured proportion. *Stat. Med.* **33** 1700–1712. MR3246689

TIAN, L. and CAI, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93** 329–342. MR2278087

WANG, L., MCMAHAN, C. S., HUDGENS, M. G. and QURESHI, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics* **72** 222–231. MR3500591

WOODWARD, M. (1999). *Epidemiology*: *Study Design and Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. MR1696292

ZHANG, Y., HUA, L. and HUANG, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scand. J. Stat.* **37** 338–354. MR2682304

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1085–1116
DOI: 10.1214/17-AOAS1038

ROBUST MIXED EFFECTS MODEL FOR CLUSTERED FAILURE TIME DATA: APPLICATION TO HUNTINGTON'S DISEASE EVENT MEASURES

BY TANYA P. GARCIA*,[1], YANYUAN MA†,[2], KAREN MARDER‡,[3] AND YUANJIA WANG‡,[4]

Texas A&M University, Pennsylvania State University*† *and Columbia University*‡

An important goal in clinical and statistical research is properly modeling the distribution for clustered failure times which have a natural intraclass dependency and are subject to censoring. We handle these challenges with a novel approach that does not impose restrictive modeling or distributional assumptions. Using a logit transformation, we relate the distribution for clustered failure times to covariates and a random, subject-specific effect. The covariates are modeled with unknown functional forms, and the random effect may depend on the covariates and have an unknown and unspecified distribution. We introduce pseudovalues to handle censoring and splines for functional covariate effects, and frame the problem into fitting an additive logistic mixed effects model. Unlike existing approaches for fitting such models, we develop semiparametric techniques that estimate the functional model parameters without specifying or estimating the random effect distribution. We show both theoretically and empirically that the resulting estimators are consistent for any choice of random effect distribution and any dependency structure between the random effect and covariates. Last, we illustrate the method's utility in an application to a Huntington's disease study where our method provides new insights into differences between motor and cognitive impairment event times in at-risk subjects.

REFERENCES

ANDERSEN, P. K. and POHAR PERME, M. (2010). Pseudo-observations in survival analysis. *Stat. Methods Med. Res.* **19** 71–99. MR2744493

BENNETT, S. (1983). Analysis of survival data by the proportional odds model. *Stat. Med.* **2** 273–277.

CHEN, M.-C. and BANDEEN-ROCHE, K. (2005). A diagnostic for association in bivariate survival models. *Lifetime Data Anal.* **11** 245–264. MR2158784

CHEN, Y., CHEN, K. and YING, Z. (2010). Analysis of multivariate failure time data using marginal proportional hazards model. *Statist. Sinica* **20** 1025–1041. MR2729851

CHEN, D. G. and LIO, Y. L. (2008). Comparative studies on frailties in survival analysis. *Comm. Statist. Simulation Comput.* **37** 1631–1646. MR2542422

CLAYTON, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65** 141–151. MR0501698

Key words and phrases. Additive model, clustered failure times, logistic mixed model, varying coefficient model, semiparametric estimator, splines.

CONGDON, P. (1994). Analyzing mortality in London: Life-tables with frailty. *Statistician* **43** 277–308.

CONNE, D., RONCHETTI, E. and VICTORIA-FESER, M.-P. (2010). Goodness of fit for generalized linear latent variables models. *J. Amer. Statist. Assoc.* **105** 1126–1134. With supplementary material available online. MR2752608

DE BOOR, C. (2001). *A Practical Guide to Splines*, revised ed. *Applied Mathematical Sciences* **27**. Springer, New York. MR1900298

DUFF, K., PAULSEN, J., MILLS, J., BEGLINGER, L. J., MOSER, D. J., SMITH, M. M., LANGBEHN, D., STOUT, J., QUELLER, S., HARRINGTON, D. L. and THE PREDICT-HD INVESTIGATORS AND COORDINATORS OF THE HUNTINGTON STUDY GROUP (2010). Mild cognitive impairment in prediagnosed Huntington disease. *Neurology* **75** 500–507.

EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425. MR0971367

GARCIA, T. P. and MA, Y. (2016). Optimal estimator for logistic model with distribution-free random intercept. *Scand. J. Stat.* **43** 156–171. MR3466999

GARCIA, T. P., MA, Y., MARDER, K. and WANG, Y. (2017). Supplement to "Robust mixed effects model for clustered failure time data: Application to Huntington's disease event measures." DOI:10.1214/17-AOAS1038SUPP.

GEERDENS, C., CLAESKENS, G. and JANSSEN, P. (2013). Goodness-of-fit tests for the frailty distribution in proportional hazards models with shared frailty. *Biostatistics* **14** 433–446.

GLIDDEN, D. V. and VITTINGHOFF, E. (2004). Modeling clustered survival data from multicenter clinical trials. *Stat. Med.* **23** 369–388.

GORFINE, M., DE-PICCIOTTO, R. and HSU, L. (2012). Conditional and marginal estimates in case-control family data—Extensions and sensitivity analyses. *J. Stat. Comput. Simul.* **82** 1449–1470. MR2971964

GOVINDARAJULU, U. S., GLICKMAN, M. E. and D'AGOSTINO, R. B. SR. (2007). Modeling frailty as a function of observed covariates. *J. Stat. Theory Pract.* **1** 117–135. MR2354620

HARPER, P. S. (1996). *Huntington's Disease*, 2nd ed. W.B. Saunders, London.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271. MR0513692

HEAGERTY, P. J. and KURLAND, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88** 973–985. MR1872214

HENDERSON, R. and OMAN, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 367–379. MR1680322

HSU, L., GORFINE, M. and MALONE, K. (2007). On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified. *Stat. Med.* **26** 4657–4678. MR2411893

HUANG, S. S., YOKOE, D. S., STELLING, J., PLACZEK, H., KULLDORFF, M., KLEINMAN, K., O'BRIEN, T. F., CALDERWOOD, M. S., VOSTOK, J., DUNN, J. and PLATT, R. (2010). Automated detection of infectious disease outbreaks in hospitals: A retrospective cohort study. *PLoS Med.* **7** e1000238.

HUBER, P., RONCHETTI, E. and VICTORIA-FESER, M.-P. (2004). Estimation of generalized linear latent variable models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 893–908. MR2102471

HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP (2010). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72** 971–983.

JOHNSON, S. G. and NARASIMHAN, B. (2013). Cubature: Adaptive multivariate integration over hypercubes. R package version 1.1-2. Available at http://CRAN.R-project.org/package=cubature.

KLEIN, J. P., VAN HOUWELINGEN, H. C., IBRAHIM, J. G. and SCHEIKE, T. H., eds. (2014). *Handbook of Survival Analysis*. CRC Press, Boca Raton, FL. MR3287588

LANGBEHN, D. R., BRINKMAN, R. R., FALUSH, D., PAULSEN, J. S. and HAYDEN, M. R. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin. Genet.* **65** 267–277.

LEE, K. J. and THOMPSON, S. G. (2008). Flexible parametric models for random-effects distributions. *Stat. Med.* **27** 418–434. MR2418453

LESAFFRE, E. and MOLENBERGHS, G. (2001). Multivariate probit analysis: A neglected procedure in medical statistics. *Stat. Med.* **10** 1391–1403.

LOGAN, B. R., NELSON, G. O. and KLEIN, J. P. (2008). Analyzing center specific outcomes in hematopoietic cell transplantation. *Lifetime Data Anal.* **14** 389–404. MR2464766

LOGAN, B. R., ZHANG, M.-J. and KLEIN, J. P. (2011). Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics* **67** 1–7. MR2898811

MA, Y. and GENTON, M. G. (2010). Explicit estimating equations for semiparametric generalized linear latent variable models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 475–495. MR2758524

MARDER, K., ZHAO, H., MYERS, R., CUDKOWICZ, M., KAYSON, E., KIEBURTZ, K., ORME, C., PAULSEN, J., PENNEY, J., SIEMERS, E., SHOULSON, I. and THE HUNTINGTON STUDY GROUP (2000). Rate of functional decline in Huntington's disease. *Neurology* **369** 452–458.

MARDER, K., LEVY, G., LOUIS, E. D., MEJIA-SANTANA, H., COTE, L., ANDREWS, H., HARRIS, J., WATERS, C., FORD, B., FRUCHT, S., FAHN, S. and OTTMAN, R. (2003). Accuracy of family history data on Parkinson's disease. *Neurology* **61** 18–23.

MURPHY, S. A., ROSSINI, A. J. and VAN DER VAART, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127

PAULSEN, J. and LONG, J. (2014). Onset of Huntington's disease: Can it be purely cognitive? *Mov. Disord.* **29** 1342–1350.

PIEPHO, H. P. and MCCULLOCH, C. E. (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. *J. Agric. Biol. Environ. Stat.* **9** 123–137.

RIPATTI, S. and PALMGREN, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56** 1016–1022. MR1806744

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959

ROSS, C. A. and TABRIZI, S. J. (2010). Huntington's disease: From molecular pathogenesis to clinical treatment. *Lancet Neurol.* **10** 83–98.

SHIH, J. H. and LOUIS, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51** 1384–1399. MR1381050

STOUT, J. C., PAULSEN, J. S., QUELLER, S., SOLOMON, A. C., WHITLOCK, K. B., CAMPBELL, J. C., CARLOZZI, N., DUFF, K., BEGLINGER, L. J., LANGBEHN, D. R., JOHNSON, S. A., BIGLAN, K. M. and AYLWARD, E. H. (2011). Neurocognitive signs in prodromal Huntington disease. *Neuropsychology* **25** 1–14.

TSIATIS, A. A. and MA, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91** 835–848. MR2126036

WOOD, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 495–518. MR2420412

YING, Z. and WEI, L. J. (1994). The Kaplan–Meier estimate for dependent failure time observations. *J. Multivariate Anal.* **50** 17–29. MR1292605

ZENG, D., LIN, D. Y. and YIN, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *J. Amer. Statist. Assoc.* **100** 470–483. MR2160551

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1117–1145
DOI: 10.1214/17-AOAS1039

VARIABLE SELECTION FOR A CATEGORICAL VARYING-COEFFICIENT MODEL WITH IDENTIFICATIONS FOR DETERMINANTS OF BODY MASS INDEX[1]

BY JITI GAO*, BIN PENG† ZHAO REN‡ AND XIAOHUI ZHANG§

Monash University, University of Bath†, University of Pittsburgh‡ and University of Exeter§*

Obesity has become one of the major public health issues during the last three decades. A considerable number of determinants have been proposed for body mass index (BMI) by a large range of studies from multiple disciplines. In addition, it is well documented that impacts of these determinants are varying across demographic groups. However, little is known about the relative importance of these potential determinants and the varying impacts of all relatively important determinants. Using the shrinkage estimation technique, we propose a variable selection procedure for the categorical varying-coefficient model. We present a simulation study to exam performance of our method in different scenarios. We further apply the proposed method to examine the impacts of a large number of potential determinants on BMI using data from the 2013 National Health Interview Survey in the United States. By our method, the relevant determinants of BMI are identified through the variable selection procedure; and their varying impacts across demographic groups are quantified through the post-selection estimation.

REFERENCES

AITCHISON, J. and AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420. MR0443222

ALI, S. M. and LINDSTRÖM, M. (2006). Socioeconomic, psychosocial, behavioural, and psychological determinants of BMI among young women: Differing patterns for underweight and overweight/obesity. *Eur. J. Public Health* **16** 324–330.

BÜHLMANN, P. and MANDOZZI, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Comput. Statist.* **29** 407–430. MR3261821

CAREY, M., SMALL, H., YOONG, S. L., BOYES, A., BISQUERA, A. and SANSON-FISHER, R. (2014). Prevalence of comorbid depression and obesity in general practice: A cross-sectional survey. *Br. J. Gen. Pract.* **64** e122–e127.

CAWLEY, J. (2011). *The Oxford Handbook of the Social Science of Obesity*. Oxford Univ. Press, Oxford.

CAWLEY, J. and SCHOLDER, S. V. H. K. (2013). The demand for cigarettes as derived from the demand for weight control. Technical Report, National Bureau of Economic Research.

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.* **10** 596–617. MR3528353

COHEN, A. K., RAI, M., REHKOPF, D. H. and ABRAMS, B. (2013). Educational attainment and obesity: A systematic review. *Obes. Rev.* **14** 989–1005.

Key words and phrases. Body mass index, obesity, optimal variable selection, varying-coefficient regression.

COLDITZ, G. A., GIOVANNUCCI, E., RIMM, E. B., STAMPFER, M. J., ROSNER, B., SPEIZER, F. E., GORDIS, E. and WILLETT, W. C. (1991). Alcohol intake in relation to diet and obesity in women and men. *Am. J. Clin. Nutr.* **54** 49–55.

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840

FAITH, M. S., BUTRYN, M., WADDEN, T. A., FABRICATORE, A., NGUYEN, A. M. and HEYMSFIELD, S. B. (2011). Evidence for prospective associations among depression and obesity in population-based studies. *Obes. Rev.* **12** e438–e453.

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

FONTAINE, K. R., REDDEN, D. T., WANG, C., WESTFALL, A. O. and ALLISON, D. B. (2003). Years of life lost due to obesity. *J. Amer. Medical Assoc.* **289** 187–193.

GALANI, C. and SCHNEIDER, H. (2007). Prevention and treatment of obesity with lifestyle interventions: Review and meta-analysis. *Int. J. Public Health* **52** 348–359.

GAO, J., PENG, B., REN, Z. and ZHANG, X. (2017). Supplement to "Variable selection for a categorical varying-coefficient model with identifications for determinants of body mass index." DOI:10.1214/17-AOAS1039SUPP.

GERTHEISS, J. and TUTZ, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat.* **4** 2150–2180. MR2829951

HALL, P., LI, Q. and RACINE, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Rev. Econ. Stat.* **89** 784–789.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HILL, J. O. and PETERS, J. C. (1998). Environmental contributions to the obesity epidemic. *Science* **280** 1371–1374.

HUANG, J., MA, S., XIE, H. and ZHANG, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

LI, Q., OUYANG, D. and RACINE, J. S. (2013). Categorical semiparametric varying-coefficient models. *J. Appl. Econometrics* **28** 551–579. MR3064528

LI, Q. and RACINE, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory* **26** 1607–1637. MR2738011

LIPOWICZ, A., GRONKIEWICZ, S. and MALINA, R. M. (2002). Body mass index, overweight and obesity in married and never married men and women in Poland. *Am. J. Human Biol.* **14** 468–475.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

MA, S., CARROLL, R. J., LIANG, H. and XU, S. (2015). Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates. *Ann. Statist.* **43** 2102–2131. MR3375878

OZA-FRANK, R. and CUNNINGHAM, S. A. (2010). The weight of US residence among immigrants: A systematic review. *Obesity Reviews* **11** 271–280.

REHKOPF, D. H., LARAIA, B. A., SEGAL, M., BRAITHWAITE, D. and EPEL, L. (2011). The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. *Int. J. Pediatr. Obes.* **6** 233–242.

SOBAL, J., RAUSCHENBACH, B. S. and FRONGILLO, E. A. (1992). Marital status, fatness and obesity. *Soc. Sci. Med.* **35** 915–923.

STICE, E., SHAW, H. and MARTI, C. N. (2006). A meta-analytic review of obesity prevention programs for children and adolescents: The skinny on interventions that work. *Psychol. Bull.* **132** 667–691.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

VON KRIES, R., TOSCHKE, A. M., KOLETZKO, B. and SLIKKER, W. (2002). Maternal smoking during pregnancy and childhood obesity. *Am. J. Epidemiol.* **156** 954–961.

WANG, H. and LENG, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102** 1039–1048. MR2411663

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

WANG, H. and XIA, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104** 747–757. MR2541592

WHO (2015). Obesity and overweight Fact Sheet No. 311, Working paper. Available at http://www.who.int/mediacentre/factsheets/fs311/en/.

YU, Y. (2012). Educational differences in obesity in the United States: A closer look at the trends. *Obes.* **20** 904–908.

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

ZENG, W., EISENBERG, D. T., JOVEL, K. R., UNDURRAGA, E. A., NYBERG, C., TANNER, S., REYES-GARCÍA, V., LEONARD, W. R., CASTANO, J., HUANCA, T. et al. (2013). Adult obesity: Panel study from native Amazonians. *Econ. Hum. Biol.* **11** 227–235.

ZHANG, Q. and WANG, Y. (2004). Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Soc. Sci. Med.* **58** 1171–1180.

ZHAO, W., ZHANG, R. and LIU, J. (2014). Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Comput. Statist. Data Anal.* **79** 44–62. MR3227986

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1146–1168
DOI: 10.1214/17-AOAS1040

LATERAL TRANSFER IN STOCHASTIC DOLLO MODELS

BY LUKE J. KELLY[1] AND GEOFF K. NICHOLLS

University of Oxford

Lateral transfer, a process whereby species exchange evolutionary traits through nonancestral relationships, is a frequent source of model misspecification in phylogenetic inference. Lateral transfer obscures the phylogenetic signal in the data as the histories of affected traits are mosaics of the overall phylogeny. We control for the effect of lateral transfer in a Stochastic Dollo model and a Bayesian setting. Our likelihood is highly intractable, as the parameters are the solution of a sequence of large systems of differential equations representing the expected evolution of traits along a tree. We illustrate our method on a data set of lexical traits in Eastern Polynesian languages, and obtain an improved fit over the corresponding model without lateral transfer.

REFERENCES

ABBY, S. S., TANNIER, E., GOUY, M. and DAUBIN, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinform.* **11** 324.

ALEKSEYENKO, A. V., LEE, C. J. and SUCHARD, M. A. (2008). Wagner and Dollo: A stochastic duet by composing two parsimonious solos. *Syst. Biol.* **57** 772–784.

BEIKO, R. G. and HAMILTON, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol.* **6** 15.

BOUCHARD-CÔTÉ, A. and JORDAN, M. I. (2013). Evolutionary inference via the Poisson Indel Process. *Proc. Natl. Acad. Sci. USA* **110** 1160–1166.

BOUCKAERT, R. and HELED, J. (2014). DensiTree 2: Seeing trees through the forest. *BioRxiv*.

BOUCKAERT, R., LEMEY, P., DUNN, M., GREENHILL, S. J., ALEKSEYENKO, A. V., DRUMMOND, A. J., GRAY, R. D., SUCHARD, M. A. and ATKINSON, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science* **337** 957–960.

BOUCKAERT, R., HELED, J., KÜHNERT, D., VAUGHAN, T., WU, C.-H., XIE, D., SUCHARD, M. A., RAMBAUT, A. and DRUMMOND, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10** 1–6.

CHANG, W., CATHCART, C., HALL, D. and GARRETT, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91** 194–244.

CONTE, E. and MOLLE, G. (2014). Reinvestigating a key site for Polynesian prehistory: New results from the Hane dune site, Ua Huka (Marquesas). *Archaeol. Ocean.* **49** 121–136.

CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** 969–991. MR3371344

DAUBIN, V., GOUY, M. and PERRIÈRE, G. (2002). A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res.* **12** 1080–1090.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. and RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29** 1969–1973.

Key words and phrases. Bayesian phylogenetics, lateral trait transfer, Stochastic Dollo model.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.

GEYER, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

GRAY, R. D. and ATKINSON, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426** 435–439.

GRAY, R. D., BRYANT, D. and GREENHILL, S. J. (2010). On the shape and fabric of human history. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 3923–3933.

GRAY, R. D., DRUMMOND, A. J. and GREENHILL, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323** 479–483.

GREENHILL, S. J., BLUST, R. and GRAY, R. D. (2008). The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evol. Bioinform.* **4** 271–283.

GREENHILL, S. J., CURRIE, T. E. and GRAY, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. Lond., B Biol. Sci.* **276** 2299–2306.

HELED, J. and DRUMMOND, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Syst. Biol.* **61** 138–149.

HUSON, D. H. and BRYANT, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23** 254–267.

HUSON, D. H. and STEEL, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics* **20** 2044–2049.

JOFRÉ, P., DAS, P., BERTRANPETIT, J. and FOLEY, R. (2017). Cosmic phylogeny: Reconstructing the chemical history of the solar neighbourhood with an evolutionary tree. *Mon. Not. R. Astron. Soc.* **467** 1140–1153.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KELLY, L. J. (2016). A Stochastic Dollo model for lateral transfer. Ph.D. thesis, Univ. Oxford.

KELLY, L. J. and NICHOLLS, G. K. (2017). Supplement to "Lateral transfer in Stochastic Dollo models." DOI:10.1214/17-AOAS1040SUPP.

KINGMAN, J. F. C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. The Clarendon Press, Oxford. MR1207584

KITCHEN, A., EHRET, C., ASSEFA, S. and MULLIGAN, C. J. (2009). Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. Lond., B Biol. Sci.* **276** 2703–2710.

KUBATKO, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.* **58** 478–488.

LATHROP, G. M. (1982). Evolutionary trees and admixture: Phylogenetic inference when some populations are hybridized. *Ann. Hum. Genet.* **46** 245–255. MR0673807

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MARCK, J. C. (2000). *Topics in Polynesian Language and Culture History* **504**. Pacific Linguistics, Canberra.

MCPHERSON, A., ROTH, A., LAKS, E., MASUD, T., BASHASHATI, A., ZHANG, A. W., HA, G., BIELE, J., YAP, D., WAN, A., PRENTICE, L. M., KHATTRA, J., SMITH, M. A., NIELSEN, C. B., MULLALY, S. C., KALLOGER, S., KARNEZIS, A., SHUMANSKY, K., SIU, C., ROSNER, J., CHAN, H. L., HO, J., MELNYK, N., SENZ, J., YANG, W., MOORE, R., MUNGALL, A. J., MARRA, M. A., BOUCHARD-CÔTÉ, A., GILKS, C. B., HUNTSMAN, D. G., MCALPINE, J. N., APARICIO, S. and SHAH, S. P. (2016). Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48** 758–767.

NICHOLLS, G. K. and GRAY, R. D. (2008). Dated ancenstral trees from binary trait data and their application to the diversification of languages. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 545–566. MR2420414

NICHOLLS, G. K. and RYDER, R. J. (2011). Phylogenetic models for Semitic vocabulary. In *Proceedings of the International Workshop on Statistical Modelling* (D. Conesa, A. Forte, A. López-Quílez and F. Muñoz, eds.) 431–436.

NICHOLLS, G. K., RYDER, R. J. and WELCH, D. (2013). TraitLab: A MatLab package for fitting and simulating binary trait-like data.

OLDMAN, J., WU, T., VAN IERSEL, L. and MOULTON, V. (2016). TriLoNet: Piecing together small networks to reconstruct reticulate evolutionary histories. *Mol. Biol. Evol.* **33** 2151–2162.

PATTERSON, N., MOORJANI, P., LUO, Y., MALLICK, S., ROHLAND, N., ZHAN, Y., GENSCHORECK, T., WEBSTER, T. and REICH, D. (2012). Ancient admixture in human history. *Genetics* **192** 1065–1093.

PICKRELL, J. K. and PRITCHARD, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8** e1002967.

RANNALA, B. and YANG, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164** 1645–1656.

ROCH, S. and SNIR, S. (2013). Recovering the treelike trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis. *J. Comput. Biol.* **20** 93–112. MR3021672

RYDER, R. J. and NICHOLLS, G. K. (2011). Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 71–92. MR2758570

SKELTON, C. (2008). Methods of using phylogenetic systematics to reconstruct the history of the Linear B script. *Archaeometry* **50** 158–176.

SPRIGGS, M. and ANDERSON, A. (1993). Late colonization of East Polynesia. *Antiquity* **67** 200–217.

SZÖLLOSI, G. J., BOUSSAU, B., ABBY, S. S., TANNIER, E. and DAUBIN, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. USA* **109** 17513–17518.

SZÖLLŐSI, G. J., TANNIER, E., LARTILLOT, N. and DAUBIN, V. (2013). Lateral gene transfer from the dead. *Syst. Biol.* **62** 386–397.

TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.

VEERAMAH, K. R., WOERNER, A. E., JOHNSTONE, L., GUT, I., GUT, M., MARQUES-BONET, T., CARBONE, L., WALL, J. D. and HAMMER, M. F. (2015). Examining phylogenetic relationships among Gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. *Genetics* **200** 295–308.

WALWORTH, M. (2014). Eastern Polynesian: The linguistic evidence revisited. *Ocean. Linguist.* **53** 256–272.

WEN, D., YU, Y. and NAKHLEH, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet.* **12** e1006006.

WILMSHURST, J. M., HUNT, T. L., LIPO, C. P. and ANDERSON, A. J. (2011). High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc. Natl. Acad. Sci. USA* **108** 1815–1820.

The Annals of Applied Statistics
2017, Vol. 11, No. 2, 1169–1192
DOI: 10.1214/17-AOAS1043

ALLELE-SPECIFIC COPY NUMBER ESTIMATION BY WHOLE EXOME SEQUENCING

BY HAO CHEN*, YUCHAO JIANG†, KARA N. MAXWELL†,1, KATHERINE L. NATHANSON†,2 AND NANCY ZHANG†,3

University of California, Davis and University of Pennsylvania†*

Whole exome sequencing is currently a technology of choice in large-scale cancer genomics studies, where the priority is to identify cancer-associated variants in coding regions. We describe a method for estimating allele-specific copy number using whole exome sequencing data from tumor and matched normal.

REFERENCES

AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J. et al. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43** 11.10.1–11.10.33.

BENJAMINI, Y. and SPEED, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40** e72.

CHEN, M., GUNEL, M. and ZHAO, H. (2013). SomatiCA: Identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. *PLoS ONE* **8** e78143.

CHEN, H., XING, H. and ZHANG, N. R. (2011). Estimation of parent specific DNA copy number in tumors using high-density genotyping arrays. *PLoS Comput. Biol.* **7** e1001060. MR2776334

CHEN, H., BELL, J. M., ZAVALA, N. A., JI, H. P. and ZHANG, N. R. (2014). Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res.* **43** e23.

FAVERO, F., JOSHI, T., MARQUARD, A. M., BIRKBAK, N. J., KRZYSTANEK, M., LI, Q., SZALLASI, Z. and EKLUND, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26** 64–70.

FLICEK, P. and BIRNEY, E. (2009). Sense from sequence reads: Methods for alignment and assembly. *Nat. Methods* **6** S6–S12.

FROMER, M., MORAN, J. L., CHAMBERT, K., BANKS, E., BERGEN, S. E., RUDERFER, D. M., HANDSAKER, R. E., MCCARROLL, S. A., O'DONOVAN, M. C., OWEN, M. J. et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91** 597–607.

JIANG, Y., OLDRIDGE, D. A., DISKIN, S. J. and ZHANG, N. R. (2015). CODEX: A normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res.* **43** e39.

KING, M.-C., MARKS, J. H., MANDELL, J. B. et al. (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302** 643–646.

KRUMM, N., SUDMANT, P. H., KO, A., O'ROAK, B. J., MALIG, M., COE, B. P., QUINLAN, A. R., NICKERSON, D. A., EICHLER, E. E., PROJECT, N. E. S. et al. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22** 1525–1532.

Key words and phrases. Allele-specific copy number, whole exome sequencing, tumor-normal pair.

LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R. and PARK, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21** 3763–3770.

MAGI, A., TATTINI, L., CIFOLA, I., D'AURIZIO, R., BENELLI, M., MANGANO, E., BATTAGLIA, C., BONORA, E., KURG, A., SERI, M. et al. (2013). EXCAVATOR: Detecting copy number variants from whole-exome sequencing data. *Genome Biol*. **14** R120.

MAXWELL, K., SLOOVER, D. D., WUBBENHORST, B., WENZ, B., JIANG, Y., CHEN, H., LUNCEFORD, N., D'ANDREA, K., EMERY, L., MORRISSETTE, J., DABER, R., MITRA, N., ZHANG, N., FELDMAN, M., DOMCHEK, S. and NATHANSON, K. (2016). Diverse mechanisms of tumor evolution in germline BRCA1/2 carriers. Working paper.

MAYRHOFER, M., DILORENZO, S., ISAKSSON, A. et al. (2013). Patchwork: Allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol*. **14** R24.

MEDVEDEV, P., STANCIU, M. and BRUDNO, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6** S13–S20.

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.

OLSHEN, A. B., BENGTSSON, H., NEUVIAL, P., SPELLMAN, P. T., OLSHEN, R. A. and SESHAN, V. E. (2011). Parent-specific copy number in paired tumor–normal studies using circular binary segmentation. *Bioinformatics* **27** 2038–2046.

PEPKE, S., WOLD, B. and MORTAZAVI, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6** S22–S32.

VENKATRAMAN, E. and OLSHEN, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23** 657–663.

WILLENBROCK, H. and FRIDLYAND, J. (2005). A comparison study: Applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21** 4084–4091.

ZHANG, N. R. (2005). Change-point detection and sequence alignment: Statistical problems of genomics. PhD thesis, Stanford University.

ZHANG, N. R. (2010). DNA copy number profiling in normal and tumor genomes. In *Frontiers in Computational and Systems Biology* 259–281. Springer, London.

ZHANG, Z., LANGE, K. and SABATTI, C. (2012). Reconstructing DNA copy number by joint segmentation of multiple sequences. *BMC Bioinform*. **13** 205.

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

ZHANG, N. R. and SIEGMUND, D. O. (2012). Model selection for high-dimensional, multi-sequence change-point problems. *Statist. Sinica* **22** 1507–1538. MR3027097