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 Vol. 13, No. 1, pp. 1–681 March 2019

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

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

President-Elect: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Past President: Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

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

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

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

IMS PUBLICATIONS

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

The Annals of Applied Statistics. *Editor-In-Chief*: Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

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

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

Statistical Science. *Editor*: Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

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

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 13, Number 1, March 2019. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to *The Annals of Applied Statistics*, Institute of Mathematical Statistics, Dues and Subscriptions Office, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 1–33
https://doi.org/10.1214/18-AOAS1185

MULTILAYER KNOCKOFF FILTER: CONTROLLED VARIABLE SELECTION AT MULTIPLE RESOLUTIONS

BY EUGENE KATSEVICH[1] AND CHIARA SABATTI[2]

Stanford University

We tackle the problem of selecting from among a large number of variables those that are "important" for an outcome. We consider situations where groups of variables are also of interest. For example, each variable might be a genetic polymorphism, and we might want to study how a trait depends on variability in genes, segments of DNA that typically contain multiple such polymorphisms. In this context, to discover that a variable is relevant for the outcome implies discovering that the larger entity it represents is also important. To guarantee meaningful results with high chance of replicability, we suggest controlling the rate of false discoveries for findings at the level of individual variables and at the level of groups. Building on the knockoff construction of Barber and Candès [*Ann. Statist.* **43** (2015) 2055–2085] and the multilayer testing framework of Barber and Ramdas [*J. Roy. Statist. Soc. Ser. B* **79** (2017) 1247–1268], we introduce the multilayer knockoff filter (MKF). We prove that MKF simultaneously controls the FDR at each resolution and use simulations to show that it incurs little power loss compared to methods that provide guarantees only for the discoveries of individual variables. We apply MKF to analyze a genetic dataset and find that it successfully reduces the number of false gene discoveries without a significant reduction in power.

REFERENCES

ABRAHAM, G., KOWALCZYK, A., ZOBEL, J. and INOUYE, M. (2012). SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinform.* **13** Art. ID 88.

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876

BARBER, R. F. and RAMDAS, A. (2017). The p-filter: Multilayer false discovery rate control for grouped hypotheses. *J. Roy. Statist. Soc. Ser. B* **79** 1247–1268. MR3689317

BENJAMINI, Y. and BOGOMOLOV, M. (2014). Selective inference on multiple families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **76** 297–318. MR3153943

BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

Key words and phrases. Variable selection, false discovery rate (FDR), group FDR, knockoff filter, p-filter, genomewide association study (GWAS), multiresolution.

BOGOMOLOV, M., PETERSON, C. B., BENJAMINI, Y. and SABATTI, C. (2017). Testing hypotheses on a tree: New error rates and controlling strategies. Preprint. Available at arXiv:1705.07529.

BRZYSKI, D., PETERSON, C. B., SOBCZYK, P., CANDÈS, E. J., BOGDAN, M. and SABATTI, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics* **205** 61–75.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'Model-X' knockoffs for high-dimensional controlled variable selection. *J. Roy. Statist. Soc. Ser. B* **80** 551–577.

DAI, R. and BARBER, R. F. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In *Proceedings of the* 33*rd International Conference on Machine Learning* (*ICML'*16). **48** 1851–1859. Available at http://proceedings.mlr.press/v48/daia16.html.

ERNST, J. and KELLIS, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **9** 215–216.

FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

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

GTEx CONSORTIUM et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660.

HELLER, R., CHATTERJEE, N., KRIEGER, A. and SHI, J. (2018). Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *J. Amer. Statist. Assoc.* **113** 1770–1783.

JALALI, A., SANGHAVI, S., RUAN, C. and RAVIKUMAR, P. K. (2010). A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems* 964–972.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KATSEVICH, E. and RAMDAS, A. (2018). Towards "simultaneous selective inference": Post-hoc bounds on the false discovery proportion. Preprint. Available at arXiv:1803.06790.

KATSEVICH, E. and SABATTI, C. (2019). Supplement to "Multilayer knockoff filter: Controlled variable selection at multiple resolutions." DOI:10.1214/18-AOAS1185SUPP.

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

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

LI, A. and BARBER, R. F. (2016). Multiple testing with the structure adaptive Benjamini–Hochberg algorithm. Preprint. Available at arXiv:1606.07926.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

MARKOVIC, J., XIA, L. and TAYLOR, J. (2017). Adaptive p-values after cross-validation. Preprint. Available at arXiv:1703.06559.

NEGAHBAN, S., YU, B., WAINWRIGHT, M. J. and RAVIKUMAR, P. K. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems* 1348–1356.

PETERSON, C. B., BOGOMOLOV, M., BENJAMINI, Y. and SABATTI, C. (2016). Many penotypes without many false discoveries: Error controlling strategies for multitrait association studies. *Genet. Epidemiol.* **40** 45–56.

POLDRACK, R. A. (2007). Region of interest analysis for fMRI. *Social Cogn. Affective Neurosci.* **2** 67–70.

RAMDAS, A., BARBER, R. F., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). A unified treatment of multiple testing with prior knowledge. Preprint. Available at arXiv:1703.06222.

RAO, N., COX, C., NOWAK, R. and ROGERS, T. T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In *Advances in Neural Information Processing Systems* 2202–2210.

SANKARAN, K. and HOLMES, S. (2014). structSSI: Simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.* **59** 1–21.

SANTORICO, S. A. and HENDRICKS, A. E. (2016). Progress in methods for rare variant association. *BMC Genet.* **17**(Suppl. 2) Art. ID 6.

SERVICE, S. K., TESLOVICH, T. M., FUCHSBERGER, C., RAMENSKY, V., YAJNIK, P., KOBOLDT, D. C., LARSON, D. E., ZHANG, Q., LIN, L., et al. (2014). Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.* **10** Art. ID e1004147.

SESIA, M., SABATTI, C. and CANDÈS, E. (2019). Gene hunting with knockoffs for hidden Markov models. *Biometrika*. **106** 1–18.

SIEGMUND, D. O., ZHANG, N. R. and YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245.

STELL, L. and SABATTI, C. (2016). Genetic variant selection: Learning across traits and sites. *Genetics* **202** 439–455.

TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123

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

WANG, K., LI, M. and BUCAN, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81** 1278–1283.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

XING, E. P., CURTIS, R. E., SCHOENHERR, G., LEE, S., YIN, J., PUNIYANI, K., WU, W. and KINNAIRD, P. (2014). GWAS in a box: Statistical and visual analytics of structured associations via GenAMap. *PLoS ONE* **9** Art. ID e97524.

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Amer. Statist. Assoc.* **103** 309–316. MR2420235

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67. MR2212574

ZHOU, H., SEHL, M. E., SINSHEIMER, J. S. and LANGE, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26** 2375–2382.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 34–59
https://doi.org/10.1214/18-AOAS1183

GROUND-LEVEL OZONE: EVIDENCE OF INCREASING SERIAL DEPENDENCE IN THE EXTREMES

BY DEBBIE J. DUPUIS[1] AND LUCA TRAPIN

HEC Montréal and Università Cattolica del Sacro Cuore

As exposure to successive episodes of high ground-level ozone concentrations can result in larger changes in respiratory function than occasional exposure buffered by lengthy recovery periods, the analysis of extreme values in a series of ozone concentrations requires careful consideration of not only the levels of the extremes but also of any dependence appearing in the extremes of the series. Increased dependence represents increased health risks and it is thus important to detect any changes in the temporal dependence of extreme values. In this paper we establish the first test for a change point in the extremal dependence of a stationary time series. The test is flexible, easy to use and can be extended along several lines. The asymptotic distributions of our estimators and our test are established. A large simulation study verifies the good finite sample properties. The test allows us to show that there has been a significant increase in the serial dependence of the extreme levels of ground-level ozone concentrations in Bloomsbury (UK) in recent years.

REFERENCES

ACKERMANN, R., HUGHES, G., HANRAHAN, D., SOMANI, A., AGGARWAL, S., FITZGERALD, A., DIXON, J., KUNTE, A., LOVEI, M. and LVOVSKY, K. (1999). *Pollution Prevention and Abatement Handbook* 1998: *Toward Cleaner Production*. World Bank Group, Washington, DC. Available at http://documents.worldbank.org/curated/en/758631468314701365/Pollution-prevention-and-abatement-handbook-1998-toward-cleaner-production.

AIR QUALITY EXPERT GROUP (2009). Ozone in the United Kingdom. Available at http://www.defra.gov.uk/environment/airquality/aqeg.

ANDREWS, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61** 821–856. MR1231678

BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121

BARNDORFF-NIELSEN, O. E., BENTH, F. E. and VERAART, A. E. D. (2011). Ambit processes and stochastic partial differential equations. In *Advanced Mathematical Methods for Finance* 35–74. Springer, Heidelberg. MR2752540

BORTOT, P. and GAETAN, C. (2014). A latent process model for temporal extremes. *Scand. J. Stat.* **41** 606–621. MR3249419

BORTOT, P. and GAETAN, C. (2016). Latent process modelling of threshold exceedances in hourly rainfall series. *J. Agric. Biol. Environ. Stat.* **21** 531–547. MR3542085

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer, London. MR1932132

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633

Key words and phrases. Threshold exceedances, hierarchical models, trawl process, change point.

DAVIS, R. A. and MIKOSCH, T. (2009). The extremogram: A correlogram for extreme events. *Bernoulli* **15** 977–1009. MR2597580

DAVIS, R. A. and YAU, C. Y. (2011). Comments on pairwise likelihood in time series models. *Statist. Sinica* **21** 255–277. MR2796862

DIERCKX, G. and TEUGELS, J. L. (2010). Change point analysis of extreme values. *Environmetrics* **21** 661–686. MR2838438

DUPUIS, D. J. (2005). Ozone concentrations: A robust analysis of multivariate extremes. *Technometrics* **47** 191–201. MR2188080

DUPUIS, D. J. (2012). Modeling waves of extreme temperature: The changing tails of four cities. *J. Amer. Statist. Assoc.* **107** 24–39. MR2949339

DUPUIS, D. J., SUN, Y. and WANG, H. J. (2015). Detecting change-points in extremes. *Stat. Interface* **8** 19–31. MR3320386

EASTOE, E. F. and TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 25–45. MR2662232

ENVIRONMENTAL RESEARCH GROUP, KING'S COLLEGE LONDON (2015). *London Air Quality Network Database*. King's College, London. Available at http://www.londonair.org.uk

FERRO, C. A. T. and SEGERS, J. (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 545–556. MR1983763

FIORE, A. M., NAIK, V., SPRACKLEN, D. V., STEINER, A., UNGER, N., PRATHER, M., BERGMANN, D., CAMERON-SMITH, P. J., CIONNI, I. et al. (2012). Global air quality and climate. *Chem. Soc. Rev.* **41** 6663–6683.

GILLELAND, E. and KATZ, R. W. (2016). extRemes 2.0: An extreme value analysis package in R. *J. Stat. Softw.* **72** 1–39.

HUSER, R. and DAVISON, A. C. (2014). Space-time modelling of extreme events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 439–461. MR3164873

JACOB, D. J. and WINNER, D. A. (2009). Effect of climate change on air quality. *Atmos. Environ.* **43** 51–63.

KIM, M. and LEE, S. (2009). Test for tail index change in stationary time series with Pareto-type marginal distribution. *Bernoulli* **15** 325–356. MR2543865

LEADBETTER, M. R. (1983). Extremes and local dependence in stationary sequences. *Z. Wahrsch. Verw. Gebiete* **65** 291–306. MR0722133

LEDFORD, A. W. and TAWN, J. A. (2003). Diagnostics for dependence within time series extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 521–543. MR1983762

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864

NOVEN, R. C., VERAART, A. E. and GANDY, A. (2015a). A latent trawl process model for extreme values. Available at http://arxiv.org/abs/1511.08190.

NOVEN, R. C., VERAART, A. E. and GANDY, A. (2017b). A latent trawl process model for extreme values. Available at http://arxiv.org/abs/1511.08190.

PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131. MR0423667

POPE, R. J. et al. (2016). The impact of synoptic weather on UK surface ozone and implications for premature mortality. *Environ. Res. Lett.* **11** 124004.

QUINTOS, C., FAN, Z. and PHILLIPS, P. C. B. (2001). Structural change tests in tail behaviour and the Asian crisis. *Rev. Econ. Stud.* **68** 633–663. MR1855475

SCHELL, J. L. and PRATHER, M. J. (2017). Co-occurrence of extremes in surface ozone, particulate matter, and temperature over eastern North America. *Proc. Natl. Acad. Sci. USA* **114** 2854–2859.

SHEN, L., MICKLEY, L. J. and GILLELAND, E. (2016). Impact of increasing heat waves on US ozone episodes in the 2050s: Results from a multimodel analysis using extreme value theory. *Geophys. Res. Lett.* **43** 4017–4025.

SHEPHARD, N. and YANG, J. J. (2017). Continuous time analysis of fleeting discrete price moves. *J. Amer. Statist. Assoc.* **112** 1090–1106. MR3735362

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201

SMITH, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statist. Sci.* **4** 367–393. MR1041763

WHO (WORLD HEALTH ORGANIZATION) (1987). *Air Quality Guidelines for Europe*. WHO Regional Office for Europe, Copenhagen.

WHO (WORLD HEALTH ORGANIZATION) (2000). *Air Quality Guidelines for Europe*, Second ed. WHO Regional Office for Europe, Copenhagen.

WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267. MR1649114

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 60–84
https://doi.org/10.1214/18-AOAS1181

GENOME-WIDE ANALYSES OF SPARSE MEDIATION EFFECTS UNDER COMPOSITE NULL HYPOTHESES

BY YEN-TSUNG HUANG[1]

Academia Sinica

A genome-wide mediation analysis is conducted to investigate whether epigenetic variations M mediate the effect of socioeconomic adversity S on adiposity Y. The mediation effect can be expressed as a product of two parameters, the S–M association and the M–Y association conditional on S. We show that the joint significance test examining the two parameters separately has smaller p-values than the normality-based or the normal product-based test for the product and is a size α test. However, under multiple tests with sparse signals, the conventional joint significance test has a conservative test size and low power within a study because of the sparsity in signals and not accounting for the composition of different null hypotheses. We develop a novel test assessing the product of two normally distributed test statistics under a composite null hypothesis, where either one parameter is zero or both are zero. We show that the null composition can be adjusted by variances of test statistics without directly estimating proportions of different nulls. Advantages of the new test are illustrated in simulation and the epigenomic study. The new test identifies four methylation loci mediating the socioeconomic effect on adiposity with the false discovery rate less than 20% while existing methods had none surviving this cut-off.

REFERENCES

AGHA, G., HOUSEMAN, E. A., KELSEY, K. T., EATON, C. B., BUKA, S. L. and LOUCKS, E. B. (2015). Adiposity is associated with DNA methylation profile in adipose tissue. *Int. J. Epidemiol.* **44** 1277–1287.

AROIAN, L. A. (1947). The probability function of the product of two normally distributed variables. *Ann. Math. Stat.* **18** 265–271. MR0021284

BARFIELD, R., SHEN, J., JUST, A. C., VOKONAS, P. S., SCHWARTZ, J., BACCARELLI, A. A., VANDERWEELE, T. J. and LIN, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genet. Epidemiol.* **41** 824–833.

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration. *J. Pers. Soc. Psychol.* **51** 1173–1182.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.* **11** 283–319. MR1445984

BORGHOL, N., SUDERMAN, M., MCARDLE, W., RACINE, A., HALLETT, M., PEMBREY, M., HERTZMAN, C., POWER, C. and SZYF, M. (2012). Associations with early-life socioeconomic position in adult DNA methylation. *Int. J. Epidemiol.* **41** 62–74.

Key words and phrases. Composite null hypothesis, epigenomics, joint significance test, mediation analysis, normal product distribution.

BULLOCK, J. G., GREEN, D. P. and HA, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *J. Pers. Soc. Psychol.* **98** 550–558.

DARMON, N. and DREWNOWSKI, A. (2008). Does social class predict diet quality? *Am. J. Clin. Nutr.* **87** 1107–1117.

DAVIS, S., DU, P., BILKE, S., TRICHE, T. J. and BOOTWALLA, M. (2015). methylumi: Handle Illumina methylation data. R package version 2.14.0.

GISKES, K., AVENDANO, M., BRUG, J. and KUNST, A. E. (2010). A systematic review of studies on socioeconomic inequalities in dietary intakes associated with weight gain and overweight/obesity conducted among European adults. *Obes. Rev.* **11** 413–429.

HUANG, Y.-T. (2015). Integrative modeling of multi-platform genomic data under the framework of mediation analysis. *Stat. Med.* **34** 162–178. MR3286246

HUANG, Y.-T (2019). Supplement to "Genome-wide analyses of sparse mediation effects under composite null hypotheses." DOI:10.1214/18-AOAS1181SUPP.

HUANG, Y.-T. and CAI, T. (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics* **72** 563–574. MR3515783

HUANG, Y.-T. and PAN, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72** 401–413. MR3515767

HUANG, Y. T., CHU, S., LOUCKS, E. B., LIN, C. L., EATON, C. B., BUKA, S. L. and KELSEY, K. T. (2016). Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood. *Epigenetics* **11** 227–236.

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. MR2741814

LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Statist. Soc. Ser. B* **67** 555–572. MR2168204

LANGE, T. and HANSEN, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology* **22** 575–581.

LOFTUS, T. M., JAWORSKY, D. E., FREHYWOT, G. L., TOWNSEND, C. A., RONNETT, G. V., LANE, M. D. and KUHAJDA, F. P. (2000). Reduced food intake and body weight in mice treated with fatty acid synthase inhibitor. *Science* **288** 2379–2381.

LOPEZ-AYALA, J. M., ORTIZ-GENGA, M., GOMEZ-MILANES, I., LOPEZ-CUENCA, D., RUIZ-ESPEJO, F., SANCHEZ-MUNOZ, J. J., OLIVA-SANDOVAL, M. J., MONSERRAT, L. and GIMENO, J. R. (2014). A mutation in the Z-line Cypher/ZASP protein is associated with arrhythmogenic right ventricular cardiomyopathy. *Clin. Genet.* **88** 172–176.

LOUCKS, E. B., HUANG, Y. T., AGHA, G., CHU, S., EATON, C. B., GILMAN, S. E., BUKA, S. L. and KELSEY, K. T. (2016). Epigenetic mediators between childhood socioeconomic disadvantage and mid-life body mass index: The New England Family Study. *Psychosom. Med.* **78** 1053–1065.

MACKINNON, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis, New York.

MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. and SHEETS, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* **7** 83–104.

NIU, T., LIU, N., ZHAO, M., XIE, G., ZHANG, L., LI, J., PEI, Y. F., SHEN, H., FU, X., HE, H., LU, S., CHEN, X. D., TAN, L. J., YANG, T. L., GUO, Y., LEO, P. J., DUNCAN, E. L., SHEN, J., GUO, Y. F., NICHOLSON, G. C., PRINCE, R. L., EISMAN, J. A., JONES, G., SAMBROOK, P. N., HU, X., DAS, P. M., TIAN, Q., ZHU, X. Z., PAPASIAN, C. J., BROWN, M. A., UITTERLINDEN, A. G., WANG, Y. P., XIANG, S. and DENG, H. W. (2015). Identification of a novel FGFRL1 microRNA target site polymorphism for bone mineral density in meta-analyses of genome-wide association studies. *Hum. Mol. Genet.* **24** 4710–4727.

PAN, W. C., WU, C. D., CHEN, M. J., HUANG, Y. T., CHEN, C. J., SU, H. J. and YANG, H. I. (2015). Fine particle pollution, alanine transaminase, and liver cancer: A Taiwanese prospective cohort study (REVEAL-HBV). *J. Natl. Cancer Inst.* **108** Art. ID djv341.

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* 411–420. Morgan Kaufmann, San Francisco, CA.

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SELCEN, D. and ENGEL, A. G. (2005). Mutations in ZASP define a novel form of muscular dystrophy in humans. *Ann. Neurol.* **57** 269–276.

SENESE, L. C., ALMEIDA, N. D., FATH, A. K., SMITH, B. T. and LOUCKS, E. B. (2009). Associations between childhood socioeconomic position and adulthood obesity. *Epidemiol. Rev.* **31** 21–51.

SOBEL, M. E. (1982). *Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models*. American Sociological Association, Washington, DC.

SONG, Y., HUANG, Y. T., SONG, Y., HEVENER, A. L., RYCKMAN, K. K., QI, L., LEBLANC, E. S., KAZLAUSKAITE, R., BRENNAN, K. M. and LIU, S. (2015). Birthweight, mediating biomarkers and the development of type 2 diabetes later in life: A prospective study of multi-ethnic women. *Diabetologia* **58** 1220–1230.

TCHETGEN TCHETGEN, E. J. (2011). On causal mediation analysis with a survival outcome. *Int. J. Biostat.* **7** Art. ID 33. MR2843528

TESCHENDORFF, A. E., MARABITA, F., LECHNER, M., BARTLETT, T., TEGNER, J., GOMEZ-CABRERO, D. and BECK, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29** 189–196.

VANDERWEELE, T. J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21** 540–551.

VANDERWEELE, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology* **22** 582–585.

VANDERWEELE, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* **24** 224–232.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* **2** 457–468. MR2576399

VANDERWEELE, T. J. and VANSTEELANDT, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.* **172** 1339–1348.

ZHANG, H., ZHENG, Y., ZHANG, Z., GAO, T., JOYCE, B., YOON, G., ZHANG, W., SCHWARTZ, J., JUST, A., COLICINO, E., VOKONAS, P., ZHAO, L., LV, J., BACCARELLI, A., HOU, L. and LIU, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32** 3150–3154.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 85–112
https://doi.org/10.1214/18-AOAS1193

COMMON AND INDIVIDUAL STRUCTURE OF BRAIN NETWORKS[1]

BY LU WANG, ZHENGWU ZHANG AND DAVID DUNSON

Central South University, University of Rochester and Duke University

This article focuses on the problem of studying shared- and individual-specific structure in replicated networks or graph-valued data. In particular, the observed data consist of n graphs, $G_i, i = 1, \ldots, n$, with each graph consisting of a collection of edges between V nodes. In brain connectomics, the graph for an individual corresponds to a set of interconnections among brain regions. Such data can be organized as a $V \times V$ binary adjacency matrix A_i for each i, with ones indicating an edge between a pair of nodes and zeros indicating no edge. When nodes have a shared meaning across replicates $i = 1, \ldots, n$, it becomes of substantial interest to study similarities and differences in the adjacency matrices. To address this problem, we propose a method to estimate a common structure and low-dimensional individual-specific deviations from replicated networks. The proposed Multiple GRAph Factorization (M-GRAF) model relies on a logistic regression mapping combined with a hierarchical eigenvalue decomposition. We develop an efficient algorithm for estimation and study basic properties of our approach. Simulation studies show excellent operating characteristics and we apply the method to human brain connectomics data.

REFERENCES

BARCH, D. M., BURGESS, G. C., HARMS, M. P., PETERSEN, S. E., SCHLAGGAR, B. L., CORBETTA, M., GLASSER, M. F., CURTISS, S., DIXIT, S., FELDT, C. et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80** 169–189.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980. DOI:10.1016/j.neuroimage.2006.01.021.

DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2014). Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Trans. Signal Process.* **62** 905–918. MR3160322

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. MR3750873

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663

Key words and phrases. Binary networks, multiple graphs, penalized logistic regression, random effects, spectral embedding.

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E., AIROLDI, E. M. et al. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

FRIEDMAN, J. HASTIE, T. J. and TIBSHIRANI, R. J. (1990). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

HEINZE, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25** 4216–4226. MR2307586

HOFF, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems* 657–664. Curran Associates, Inc., Red Hook, NY.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

KRAVITZ, D. J., SALEEM, K. S., BAKER, C. I. and MISHKIN, M. (2011). A new neural framework for visuospatial processing. *Nat. Rev., Neurosci.* **12** 217–230.

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. MR3086429

MINKA, T. P. (2003). A comparison of numerical optimizers for logistic regression. Unpublished manuscript.

MOORE, T. M., SCOTT, J. C., REISE, S. P., PORT, A. M., JACKSON, C. T., RUPAREL, K., SAVITT, A. P., GUR, R. E. and GUR, R. C. (2015). Development of an abbreviated form of the Penn Line Orientation Test using large samples and computerized adaptive test simulation. *Psychol. Assess.* **27** 955–964.

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

O'CONNOR, L., MÉDARD, M. and FEIZI, S. (2015). Clustering over logistic random dot product graphs. *Stat* **1050** 3.

PARLETT, B. N. (1998). *The Symmetric Eigenvalue Problem. Classics in Applied Mathematics* **20**. SIAM, Philadelphia, PA. Corrected reprint of the 1980 original. MR1490034

SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. MR3010899

TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In *IEEE International Conference on Data Mining* 1016–1021.

TANG, R., KETCHA, M., VOGELSTEIN, J. T., PRIEBE, C. E. and SUSSMAN, D. L. (2016). Law of large graphs. ArXiv preprint. Available at arXiv:1609.01672.

TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. MR0205395

VAN ESSEN, D. C., UGURBIL, K., AUERBACH, E., BARCH, D., BEHRENS, T., BUCHOLZ, R., CHANG, A., CHEN, L., CORBETTA, M., CURTISS, S. W. et al. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage* **62** 2222–2231.

WANG, L., ZHANG, Z. and DUNSON, D. (2019). Supplement to "Common and individual structure of brain networks." DOI:10.1214/18-AOAS1193SUPP.

WOOLFE, F., LIBERTY, E., ROKHLIN, V. and TYGERT, M. (2008). A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.* **25** 335–366. MR2455599

WU, X., LV, X.-F., ZHANG, Y.-L., WU, H.-W., CAI, P.-Q., QIU, Y.-W., ZHANG, X.-L. and JIANG, G.-H. (2015). Cortical signature of patients with HBV-related cirrhosis without overt hepatic encephalopathy: A morphometric analysis. *Front. Neuroanatom.* **9** 82.

ZEKI, S., WATSON, J., LUECK, C., FRISTON, K. J., KENNARD, C. and FRACKOWIAK, R. (1991). A direct demonstration of functional specialization in human visual cortex. *J. Neurosci.* **11** 641–649.

ZHANG, Z., DESCOTEAUX, M., ZHANG, J., GIRARD, G., CHAMBERLAND, M., DUNSON, D., SRIVASTAVA, A. and ZHU, H. (2018a). Mapping population-based structural connectomes. *NeuroImage* **172** 130–145.

ZHANG, Z., ALLEN, G., ZHU, H. and DUNSON, D. (2018b). Relationships between human brain structural connectomes and traits. *BioRxiv* 256933.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 113–131
https://doi.org/10.1214/18-AOAS1197

CLONALITY: POINT ESTIMATION[1]

BY LU TIAN*, YI LIU†, ANDREW Z. FIRE*, SCOTT D. BOYD* AND RICHARD A. OLSHEN*

Stanford University and Calico Life Sciences LLC†*

Assessments of biological complexity for populations that are of mixed species are central in many biological contexts, including microbiomes, tumor cell population structure, and immune cell populations. Here we address the problem of quantifying the population diversity in experiments where high throughput DNA sequencing is used to distinguish a large number of cell subpopulations. Our model assumes a list of clonal species and their observed frequencies in each of several replicate sequencing libraries. Though the underlying distribution of frequencies cannot be estimated well from data coming from only a small fraction of the total cell population, one can estimate well the population-level *clonality*, defined as the sum of squared underlying fractions of the respective clones, the complement of the Gini–Simpson index. Specifically, we proposed to adaptively combine multiple unbiased estimators of *clonality* derived from pairs of replicates to construct a single estimator without relying on the commonly used but restrictive multinomial assumption. The new estimator performs particularly well for replicates of unequal size. We further illustrate the proposed methods with extensive simulations and a small real data example.

REFERENCES

ALDRICH, R. J. (2010). *GCHQ*: *The Uncensored Story of Britain's Most Secret Intelligence Agency*. Harper Collins, London.

BOYD, S. D., MARSHALL, E. L., MERKER, J. D., MANIAR, J. M., ZHANG, L. N., SAHAF, B., JONES, C. D., SIMEN, B. B., HANCZARUK, B., NGUYEN, K. D., NADEAU, K. C., EGHOLM, M., MIKLOS, D. B., ZEHNDER, J. L. and FIRE, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* **1** 12a23.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43** 783–791. MR0920467

CHAO, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45** 427–438. MR1010510

EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York. MR0898653

Key words and phrases. Clonality, V(D)J rearrangements, richness, jackknife.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

KAPLINSKY, J. and ARANOUT, R. (2016). Robust estimates of overall immune-repertoire diversity from high throughput measurements on samples. *Nat. Commun.* **7** 11881. DOI:10.1038/ncomms11881.

LAYDON, D. J., BANGHAM, C. R. M. and ASQUITH, B. (2015). Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. B* **370** 20140291. DOI:10.1098/rstb.2014.0291.

MCKANE, A. G., ALONSO, D. and SOLÉ, R. V. (2004). Analytic solution of Hubbell's model of local community dynamics. *Theor. Popul. Biol.* **65** 67–73.

MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1–15. MR0391366

PARAMESWARAN, P., LIU, Y., ROSKIN, K. M., JACKSON, K. K., DIXIT, V. F., LEE, J. Y., ARTILES, K. S., ZOMPI, S., VARGAS, M. J., et al. (2013). Convergent antibody signatures in human dengue. *Cell Host Microbe* **13** 691–700.

QI, Q., LIU, Y., CHENG, Y., GLANVILLE, J., ZHANG, D., LEE, J.-Y., OLSHEN, R. A., WEYAND, C. M., BOYD, S. and GORONZY, J. J. (2014). Diversity and clonal selection in human T cell repertoire. *Proc. Natl. Acad. Sci. USA* **111** 13139–13144.

ROBBINS, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Stat.* **39** 256–257. MR0221695

SCHATZ, D. G. and JI, Y. (2011). Recombination centres and the orchestration of V (D) J recombination. *Nat. Rev., Immunol.* **11** 251–263.

TIAN, L., GREENBERG, S. A., KONG, S. S., ALTSCHULER, J., KOHANE, I. S. and PARK, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci.* **102** 13544–9.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 132–164
https://doi.org/10.1214/18-AOAS1205

PREDICTION MODELS FOR NETWORK-LINKED DATA[1]

BY TIANXI LI[*,2], ELIZAVETA LEVINA[†,3] AND JI ZHU[†,4]

University of Virginia[*] and University of Michigan[†]

Prediction algorithms typically assume the training data are independent samples, but in many modern applications samples come from individuals connected by a network. For example, in adolescent health studies of risk-taking behaviors, information on the subjects' social network is often available and plays an important role through network cohesion, the empirically observed phenomenon of friends behaving similarly. Taking cohesion into account in prediction models should allow us to improve their performance. Here we propose a network-based penalty on individual node effects to encourage similarity between predictions for linked nodes, and show that incorporating it into prediction leads to improvement over traditional models both theoretically and empirically when network cohesion is present. The penalty can be used with many loss-based prediction methods, such as regression, generalized linear models, and Cox's proportional hazard model. Applications to predicting levels of recreational activity and marijuana usage among teenagers from the AddHealth study based on both demographic covariates and friendship networks are discussed in detail and show that our approach to taking friendships into account can significantly improve predictions of behavior while providing interpretable estimates of covariate effects.

REFERENCES

ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

ASUR, S. and HUBERMAN, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology* (*WI-IAT*), 2010 *IEEE/WIC/ACM International Conference on* **1** 492–499. IEEE, New York.

BELKIN, M. and NIYOGI, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15** 1373–1396.

BELKIN, M., NIYOGI, P. and SINDHWANI, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7** 2399–2434. MR2274444

BENGIO, Y., PAIEMENT, J.-F., VINCENT, P., DELALLEAU, O., LE ROUX, N. and OUIMET, M. (2004). Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Adv. Neural Inf. Process. Syst.* **16** 177–184.

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

BINKIEWICZ, N., VOGELSTEIN, J. T. and ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104** 361–377. MR3698259

Key words and phrases. Network cohesion, prediction, regression.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993

BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454

CAI, D., HE, X. and HAN, J. (2007). Spectral regression: A unified approach for sparse subspace learning. In *Seventh IEEE International Conference on Data Mining* (*ICDM* 2007) 73–82. IEEE, New York.

CHAUDHURI, K., GRAHAM, F. C. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *COLT* **23** 35–1.

CHOI, D. (2017). Estimation of monotone treatment effects in network experiments. *J. Amer. Statist. Assoc.* **112** 1147–1155. MR3735366

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.

COHEN, M. B., KYNG, R., MILLER, G. L., PACHOCKI, J. W., PENG, R., RAO, A. B. and XU, S. C. (2014). Solving SDD linear systems in nearly $m \log^{1/2} n$ time. In *STOC'*14—*Proceedings of the* 2014 *ACM Symposium on Theory of Computing* 343–352. ACM, New York. MR3238960

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

CRESSIE, N. (1990). The origins of kriging. *Math. Geol.* **22** 239–252. MR1047810

FUJIMOTO, K. and VALENTE, T. W. (2012). Social network influences on adolescent substance use: Disentangling structural equivalence from cohesion. *Soc. Sci. Med.* **74** 1952–1960.

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

HALLAC, D., LESKOVEC, J. and BOYD, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 387–396. ACM, New York.

HARRIS, K. M. (2009). *The National Longitudinal Study of Adolescent to Adult Health* (*Add Health*), *Waves I & II*, 1994–1996; *Wave III*, 2001–2002; *Wave IV*, 2007–2009 [*Machine-Readable Data File and Documentation*]. Carolina Population Center, Univ. North Carolina at Chapel Hill, Chapel Hill.

HAYNIE, D. L. (2001). Delinquent peers revisited: Does network structure matter? *Amer. J. Sociol.* **106** 1013–1057.

HENDERSON, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9** 226–252. MR0055650

HOTHORN, T., BUEHLMANN, P., KNEIB, T., SCHMID, M. and HOFNER, B. (2018). mboost: Model-Based Boosting. R package version 2.9-0.

KIM, S., PAN, W. and SHEN, X. (2013). Network-based penalized regression with application to genomic data. *Biometrics* **69** 582–593. MR3106586

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. Springer, New York. MR2724362

KOUTIS, I., MILLER, G. L. and PENG, R. (2010). Approaching optimality for solving SDD linear systems. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 235–244. IEEE Computer Soc., Los Alamitos, CA. MR3024797

LAND, S. R. and FRIEDMAN, J. H. (1997). Variable fusion: A new adaptive signal regression method. Technical Report 656, Department of Statistics, Carnegie Mellon Univ., Pittsburgh, PA.

LE, C. M., LEVINA, E. and VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Structures Algorithms* **51** 538–561. MR3689343

LEE, L. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econometrics* **140** 333–374. MR2408910

LEE, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *J. Stat. Softw.* **55** 1–24.

LI, T., LEVINA, E. and ZHU, J. (2016). netcoh: Statistical Modeling with Network Cohesion. R package version 0.11.

LI, T., LEVINA, E. and ZHU, J. (2019). Supplement to "Prediction models for network-linked data." DOI:10.1214/18-AOAS1205SUPP.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4** 1498–1516. MR2758338

LIN, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *J. Labor Econ.* **28** 825–860.

MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60** 531–542. MR1236836

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060

MICHELL, L. and WEST, P. (1996). Peer pressure to smoke: The meaning depends on the method. *Health Educ. Res.* **11** 39–49.

NEWMAN, M. E. J. and CLAUSET, A. (2016). Structure and inference in annotated networks. *Nat. Commun.* **7** 11863.

PAN, W., XIE, B. and SHEN, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* **66** 474–484. MR2758827

PEARSON, M. and MICHELL, L. (2000). Smoke rings: Social network analysis of friendship groups, smoking and drug-taking. *Drugs Educ. Prev. Policy* **7** 21–37.

PEARSON, M. and WEST, P. (2003). Drifting smoke rings. *Connections* **25** 59–76.

PHAN, T. Q. and AIROLDI, E. M. (2015). A natural experiment of social network formation and dynamics. *Proc. Natl. Acad. Sci. USA* **112** 6595–6600.

RADUCANU, B. and DORNAIKA, F. (2012). A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognit.* **45** 2432–2444.

RAND, D. G., ARBESMAN, S. and CHRISTAKIS, N. A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl. Acad. Sci. USA* **108** 19193–19198.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347

SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Graph sparsification approaches for Laplacian smoothing. In *Proceedings of the* 19*th International Conference on Artificial Intelligence and Statistics* 1250–1259.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. MR1190470

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

SHARPNACK, J., SINGH, A. and KRISHNAMURTHY, A. (2013). Detecting activations over graphs using spanning tree wavelet bases. In *Artificial Intelligence and Statistics* 536–544.

SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

SONG, X. and ZHOU, X.-H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statist. Sinica* **18** 947–965. MR2440075

SPIELMAN, D. A. and TENG, S.-H. (2011). Spectral sparsification of graphs. *SIAM J. Comput.* **40** 981–1025. MR2825307

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

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

VAPNIK, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. MR1367965

VOGELSTEIN, J. T., RONCAL, W. G., VOGELSTEIN, R. J. and PRIEBE, C. E. (2013). Graph classification using signal-subgraphs: Applications in statistical connectomics. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1539–1551.

VURAL, E. and GUILLEMOT, C. (2016). Out-of-sample generalizations to supervised manifold learning for classification. *IEEE Trans. Image Process.* **25** 1410–1424. MR3464976

WAHBA, G. et al. (1999). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods-Support Vector Learning* **6** 69–87.

WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley, Hoboken, NJ. MR2075123

WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J. Mach. Learn. Res.* **17** Paper No. 105, 41. MR3543511

WOLF, T., SCHROTER, A., DAMIAN, D. and NGUYEN, T. (2009). Predicting build failures using social network analysis on developer communication. In *Proceedings of the* 31*st International Conference on Software Engineering* 1–11. IEEE Comput. Soc., Los Alamitos, CA.

XU, Y., DYER, J. S. and OWEN, A. B. (2010). Empirical stationary correlations for semi-supervised learning on graphs. *Ann. Appl. Stat.* **4** 589–614. MR2758641

YANG, W., SUN, C. and ZHANG, L. (2011). A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognit.* **44** 1649–1657.

ZHANG, Y., LEVINA, E. and ZHU, J. (2016). Community detection in networks with node features. *Electron. J. Stat.* **10** 3153–3178. MR3571965

ZHOU, D., HUANG, J. and SCHÖLKOPF, B. (2005). Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the* 22*nd International Conference on Machine Learning* 1036–1043. ACM, New York.

ZHOU, D., BOUSQUET, O., LAL, T. N., WESTON, J. and SCHÖLKOPF, B. (2004). Learning with local and global consistency. In *Advances in Neural Information Processing Systems* 321–328.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 165–188
https://doi.org/10.1214/18-AOAS1204

NONSTATIONARY SPATIAL PREDICTION OF SOIL ORGANIC CARBON: IMPLICATIONS FOR STOCK ASSESSMENT DECISION MAKING[1]

BY MARK D. RISSER*, CATHERINE A. CALDER[†,2], VERONICA J. BERROCAL[‡] AND CANDACE BERRETT[§]

Lawrence Berkeley National Laboratory, Ohio State University[†], University of Michigan[‡] and Brigham Young University[§]*

The Rapid Carbon Assessment (RaCA) project was conducted by the US Department of Agriculture's National Resources Conservation Service between 2010–2012 in order to provide contemporaneous measurements of soil organic carbon (SOC) across the US. Despite the broad extent of the RaCA data collection effort, direct observations of SOC are not available at the high spatial resolution needed for studying carbon storage in soil and its implications for important problems in climate science and agriculture. As a result, there is a need for predicting SOC at spatial locations not included as part of the RaCA project. In this paper, we compare spatial prediction of SOC using a subset of the RaCA data for a variety of statistical methods. We investigate the performance of methods with off-the-shelf software available (both stationary and nonstationary) as well as a novel nonstationary approach based on partitioning relevant spatially-varying covariate processes. Our new method addresses open questions regarding (1) how to partition the spatial domain for segmentation-based nonstationary methods, (2) incorporating partially observed covariates into a spatial model, and (3) accounting for uncertainty in the partitioning. In applying the various statistical methods we find that there are minimal differences in out-of-sample criteria for this particular data set, however, there are major differences in maps of uncertainty in SOC predictions. We argue that the spatially-varying measures of prediction uncertainty produced by our new approach are valuable to decision makers, as they can be used to better benchmark mechanistic models, identify target areas for soil restoration projects, and inform carbon sequestration projects.

REFERENCES

ANGERS, D. A., ARROUAYS, D., SABY, N. P. A. and WALTER, C. (2011). Estimating and mapping the carbon saturation deficit of French agricultural topsoils. *Soil Use Manage*. **27** 448–452.

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **70** 825–848. MR2523906

BATJES, N. H. (1996). Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci*. **47** 151–163. DOI:10.1111/j.1365-2389.1996.tb01386.x.

BEAUDETTE, D. E. and SKOVLIN, J. M. (2015). `soilDB`: Soil database interface. R package version 1.5-2.

Key words and phrases. Gaussian process, spatial clustering, model averaging, soil carbon, spatial regression.

BLISS, N. B., WALTMAN, S. W., WEST, L. T., NEALE, A. and MEHAFFEY, M. (2014). Distribution of soil organic carbon in the conterminous United States. In *Soil Carbon* (A. E. Hartemink and K. McSweeney, eds.). *Progress in Soil Science* 85–93. Springer, Berlin.

BONATO, V., BALADANDAYUTHAPANI, V., BROOM, B. M., SULMAN, E. P., ALDAPE, K. D. and DO, K.-A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27** 359–367.

CALDER, C. A. (2008). A dynamic process convolution approach to modeling ambient particulate matter concentrations. *Environmetrics* **19** 39–48. MR2416543

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413. MR3640196

DING, J., BASHASHATI, A., ROTH, A., OLOUMI, A., TSE, K., ZENG, T., HAFFARI, G., HIRST, M., MARRA, M. A., CONDON, A., APARICIO, S. and SHAH, S. P. (2012). Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28** 167–175.

FINLEY, A. O., BANERJEE, S. and CARLIN, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *J. Stat. Softw.* **19** 1.

FINLEY, A. O., BANERJEE, S. and GELFAND, A. E. (2013). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. Preprint. Available at arXiv:1310.8192.

FUENTES, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics* **12** 469–483.

FUGLSTAD, G.-A., SIMPSON, D., LINDGREN, F. and RUE, H. (2015). Does non-stationary spatial data always require non-stationary random fields? *Spat. Stat.* **14** 505–531. MR3431054

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284

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

GOIDTS, E. and VAN WESEMAEL, B. (2007). Regional assessment of soil organic carbon changes under agriculture in Southern Belgium, 1955–2005. *Geoderma* **141** 341–354.

GRAMACY, R. B. (2007). tgp: An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *J. Stat. Softw.* **19** 1–46.

GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statist. Assoc.* **103** 1119–1130. MR2528830

GREEN, D. P. and KERN, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* **76** 491–511.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.* **14** 382–417. MR1765176

INGEBRIGTSEN, R., LINDGREN, F. and STEINSLAND, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spat. Stat.* **8** 20–38. MR3326819

JOBBÁGY, E. G. and JACKSON, R. B. (2000). The vertical distribution of soil organic carbon and its relationship to climate and vegetation. *Ecol. Appl.* **10**.

JORDAN, A., KRÜGER, F. and LERCH, S. (2016). scoringRules: Scoring rules for parametric and simulated distribution forecasts. R package version 0.9.

KATZFUSS, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics* **24** 189–200. MR3067342

KRÜGER, F., LERCH, S., THORARINSDOTTIR, T. and GNEITING, T. (2016). Probabilistic forecasting and comparative model assessment based on Markov chain Monte Carlo output. Preprint. Available at arXiv:1608.06802.

LEISCH, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.* **11** 1–18. DOI:10.18637/jss.v011.i08.

MINASNY, B., MCBRATNEY, A. B., MENDONCA-SANTOS, M. L., ODEH, I. O. A. and GUYON, B. (2006). Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Soil Res.* **44** 233–244.

MISHRA, U., LAL, R., SLATER, B., CALHOUN, F., LIU, D. and VAN MEIRVENNE, M. (2009). Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Sci. Soc. Amer. J.* **73** 614–621.

MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. MR2816548

PACIOREK, C. J. and SCHERVISH, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17** 483–506. MR2240939

POST, W. M. and KWON, K. C. (2000). Soil carbon sequestration and land-use change: Processes and potential. *Glob. Change Biol.* **6** 317–327.

POST, W. M., EMANUEL, W. R., ZINKE, P. J. and STANGENBERGER, A. (1982). Soil carbon pools and world life zones. *Nature* **298** 156–159.

REICH, B. J., EIDSVIK, J., GUINDANI, M., NAIL, A. J. and SCHMIDT, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *Ann. Appl. Stat.* **5** 2425–2447. MR2907121

RISSER, M. D. and CALDER, C. A. (2015). Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics* **26** 284–297. MR3340964

RISSER, M. D., CALDER, C. A., BERROCAL, V. J. and BERRETT, C. (2019). Supplement to "Nonstationary spatial prediction of soil organic carbon: Implications for stock assessment decision making." DOI:10.1214/18-AOAS1204SUPP.

RYALS, R., KAISER, M., TORN, M. S., BERHE, A. A. and SILVER, W. L. (2014). Impacts of organic matter amendments on carbon and nitrogen dynamics in grassland soils. *Soil Biol. Biochem.* **68** 52–61.

RYALS, R., HARTMAN, M. D., PARTON, W. J., DELONGE, M. S. and SILVER, W. L. (2015). Long-term climate change mitigation potential with organic matter management on grasslands. *Ecol. Appl.* **25** 531–545.

SCHMIDT, A. M., GUTTORP, P. and O'HAGAN, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics* **22** 487–500. MR2843403

SIMBAHAN, G. C., DOBERMANN, A., GOOVAERTS, P., PING, J. and HADDIX, M. L. (2006). Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma* **132** 471–489.

SLEUTEL, S., DE NEVE, S., HOFMAN, G., BOECKX, P., BEHEYDT, D., VAN CLEEMPUT, O., MESTDAGH, I., LOOTENS, P., CARLIER, L., VAN CAMP, N., VERBEECK, H., VANDE WALLE, I., SAMSON, R., LUST, N. and LEMEUR, R. (2003). Carbon stock changes and carbon sequestration potential of Flemish cropland soils. *Glob. Change Biol.* **9** 1193–1203.

TODD-BROWN, K. E. O., RANDERSON, J. T., HOPKINS, F., ARORA, V., HAJIMA, T., JONES, C., SHEVLIAKOVA, E., TJIPUTRA, J., VOLODIN, E., WU, T., ZHANG, Q. and ALLISON, S. D. (2014). Changes in soil organic carbon storage predicted by Earth system models during the 21st century. *Biogeosciences* **11** 2341–2356.

VIANNA NETO, J. H., SCHMIDT, A. M. and GUTTORP, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 103–122. MR3148271

WIESMEIER, M., HÜBNER, R., SPÖRLEIN, P., GEUSS, U., HANGEN, E., REISCHL, A., SCHILLING, B., LÜTZOW, M. and KÖGEL-KNABNER, I. (2014). Carbon sequestration potential of soils in southeast Germany derived from stable soil organic carbon saturation. *Glob. Change Biol*. **20** 653–665.

WILLS, S., SEQUEIRA, C., LOECKE, T., BENHAM, E., FERGUSON, R., SCHEFFE, K. and WEST, L. (2013). Rapid Carbon Assessment (RaCA) Methodology Sampling and Initial Summary. United States Department of Agriculture, Natural Resources Conservation Service. Available online at http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_052841.pdf.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 189–220
https://doi.org/10.1214/18-AOAS1201

AN ALGORITHM FOR REMOVING SENSITIVE INFORMATION: APPLICATION TO RACE-INDEPENDENT RECIDIVISM PREDICTION[1]

BY JAMES E. JOHNDROW AND KRISTIAN LUM

Stanford University and Human Rights Data Analysis Group

Predictive modeling is increasingly being employed to assist human decision-makers. One purported advantage of replacing or augmenting human judgment with computer models in high stakes settings—such as sentencing, hiring, policing, college admissions, and parole decisions—is the perceived "neutrality" of computers. It is argued that because computer models do not hold personal prejudice, the predictions they produce will be equally free from prejudice. There is growing recognition that employing algorithms does not remove the potential for bias, and can even amplify it if the training data were generated by a process that is itself biased. In this paper, we provide a probabilistic notion of algorithmic bias. We propose a method to eliminate bias from predictive models by removing all information regarding protected variables from the data to which the models will ultimately be trained. Unlike previous work in this area, our procedure accommodates data on any measurement scale. Motivated by models currently in use in the criminal justice system that inform decisions on pre-trial release and parole, we apply our proposed method to a dataset on the criminal histories of individuals at the time of sentencing to produce "race-neutral" predictions of re-arrest. In the process, we demonstrate that a common approach to creating "race-neutral" models—omitting race as a covariate—still results in racially disparate predictions. We then demonstrate that the application of our proposed method to these data removes racial disparities from predictions with minimal impact on predictive accuracy.

REFERENCES

ADLER, P., FALK, C., FRIEDLER, S. A., RYBECK, G., SCHEIDEGGER, C., SMITH, B. and VENKATASUBRAMANIAN, S. (2016). Auditing black-box models for indirect influence. In *IEEE International Conference on Data Mining*.

ALPERT, G. P., SMITH, M. R. and DUNHAM, R. G. (2004). Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Res. Policy* **6** 43–69.

ANGWIN, J., LARSON, J., MATTU, S. and KIRCHNER, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

BAROCAS, S. and SELBST, A. D. (2016). Big data's disparate impact. *Calif. Law Rev.* **104** 671–732.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERK, R. (2016). A primer on fairness in criminal justice risk assessment. *The Criminologist* **41** 6–9.

Key words and phrases. Algorithmic fairness, criminal justice, neutral predictions, racial bias, recidivism, risk assessment, selection bias.

BERK, R., SHERMAN, L., BARNES, G., KURTZ, E. and AHLMAN, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *J. Roy. Statist. Soc. Ser. A* **172** 191–211. MR2655611

BRENNAN, T., DIETERICH, W. and EHRET, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Crim. Justice Behav.* **36** 21–40.

BRIDGES, G. S. and CRUTCHFIELD, R. D. (1988). Law, social standing and racial disparities in imprisonment. *Soc. Forces* **66** 699–724.

BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**.

CALDERS, T. and VERWER, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* **21** 277–292. MR2720507

CHOULDECHOVA, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **5** 153–163.

CUNNINGHAM, M. D. and SORENSEN, J. R. (2006). Actuarial models for assessing prison violence risk revisions and extensions of the risk assessment scale for prison (RASP). *Assessment* **13** 253–265.

DALL'AGLIO, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Ann. Sc. Norm. Super. Pisa* (3) **10** 35–74. MR0081577

DIETERICH, W., MENDOZA, C. and BRENNAN, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe.

DVOSKIN, J. A. and HEILBRUN, K. (2001). Risk assessment and release decision-making: Toward resolving the great debate. *J. Am. Acad. Psychiatry Law* **29** 6–10.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O. and ZEMEL, R. (2012). Fairness through awareness. In *Proceedings of the* 3*rd Innovations in Theoretical Computer Science Conference* 214–226. ACM, New York. MR3388391

EKISHEVA, S. and HOUDRÉ, C. (2006). Transportation distance and the central limit theorem. ArXiv preprint, Math/0607089.

FELDMAN, M., FRIEDLER, S. A., MOELLER, J., SCHEIDEGGER, C. and VENKATASUBRAMANIAN, S. (2015). Certifying and removing disparate impact. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 259–268. ACM.

FLORES, A. W., BECHTEL, K. and LOWENKAMP, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Fed. Probat.* **80** 38–46.

GLASER, J. (2014). *Suspect Race*: *Causes and Consequences of Racial Profiling*. Oxford Univ. Press, New York.

HARDT, M., PRICE, E., SREBRO, N. et al. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 3315–3323.

HOFFMAN, M., KAHN, L. B. and LI, D. (2015). Discretion in hiring. Technical report, National Bureau of Economic Research.

KAMIRAN, F. and CALDERS, T. (2009). Classifying without discriminating. In 2*nd International Conference on Computer, Control and Communication* IEEE.

KHANDANI, A. E., KIM, A. J. and LO, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *J. Bank. Financ.* **34** 2767–2787.

KLEINBERG, J., MULLAINATHAN, S. and RAGHAVAN, M. (2017). Inherent trade-offs in the fair determination of risk scores. In 8*th Innovations in Theoretical Computer Science Conference*. *LIPIcs. Leibniz Int. Proc. Inform.* **67** Art. No. 43, 23. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. Available at arXiv:1609.05807. MR3754967

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* 4066–4076.

LANGAN, P. A. (1995). The racial disparity in U.S. drug arrests. Bureau of Justice Statistics (BJS) and US Dept. Justice and Office of Justice Programs and United States of America.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Stat.* **43** 508–515. MR0298812

MITCHELL, O. and CAUDY, M. S. (2015). Examining racial disparities in drug arrests. *Justice Q.* **32** 288–313.

PHILLIPS, M. T., FERRI, R. F. and CALIGIURE, R. P. (2016). Annual report 2014. Technical report, Criminal Justice Agency.

QUINSEY, V. L., HARRIS, G. T., RICE, M. E. and CORMIER, C. A. (2006). *Violent Offenders: Appraising and Managing Risk*, 2nd ed. American Psychological Association, Washington, DC.

REITER, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. MR2113234

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542

ROMEI, A. and RUGGIERI, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.* **29** 582–638.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. Reprint of the 1987 edition [John Wiley & Sons, Inc., New York]. MR2117498

RUDOVSKY, D. (2001). Law enforcement by stereotypes and serendipity: Racial profiling and stops and searches without cause. *Univ. Pa. J. Const. Law* **3** 296.

SALVEMINI, T. (1943). Sul calcolo degli indici di concordanza tra due caratteri quantitativi. *Atti Riun. Sci. - Soc. Ital. Stat.*.

SIMOIU, C., CORBETT-DAVIES, S. and GOEL, S. (2016). Testing for racial discrimination in police searches of motor vehicles. *SSRN Electron. J.* 2811449.

TAYLOR, M. (2015). No one gets hurt: Why the future of crime may be less violent and more insidious. *Calif. Mag.* **Summer 2015**.

WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30** 377–399. MR2758870

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G. and GUMMADI, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the* 26*th International Conference on World Wide Web* 1171–1180.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 221–247
https://doi.org/10.1214/18-AOAS1182

BAYESIAN SEMIPARAMETRIC JOINT REGRESSION ANALYSIS OF RECURRENT ADVERSE EVENTS AND SURVIVAL IN ESOPHAGEAL CANCER PATIENTS

BY JUHEE LEE[*,1], PETER F. THALL[†,2] AND STEVEN H. LIN[†]

University of California, Santa Cruz[*] and
University of Texas MD Anderson Cancer Center[†]

We propose a Bayesian semiparametric joint regression model for a recurrent event process and survival time. Assuming independent latent subject frailties, we define marginal models for the recurrent event process intensity and survival distribution as functions of the subject's frailty and baseline covariates. A robust Bayesian model, called Joint-DP, is obtained by assuming a Dirichlet process for the frailty distribution. We present a simulation study that compares posterior estimates under the Joint-DP model to a Bayesian joint model with lognormal frailties, a frequentist joint model, and marginal models for either the recurrent event process or survival time. The simulations show that the Joint-DP model does a good job of correcting for treatment assignment bias, and has favorable estimation reliability and accuracy compared with the alternative models. The Joint-DP model is applied to analyze an observational dataset from esophageal cancer patients treated with chemoradiation, including the times of recurrent effusions of fluid to the heart or lungs, survival time, prognostic covariates, and radiation therapy modality.

REFERENCES

AUSTIN, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Stat. Med.* **32** 2837–2849. MR3069909

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BROWN, E. R. and IBRAHIM, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59** 221–228. MR1987388

BUSH, C. A. and MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83** 275–285.

CHAPPLE, A. G., VANNUCCI, M., THALL, P. F. and LIN, S. (2017). Bayesian variable selection for a semi-competing risks model with three hazard functions. *Comput. Statist. Data Anal.* **112** 170–185. MR3645597

CHUONG, M. D., HALLEMEIER, C. L., JABBOUR, S. K., YU, J., BADIYAN, S., MERRELL, K. W., MISHRA, M. V., LI, H., VERMA, V. and LIN, S. H. (2016). Improving outcomes for esophageal cancer using proton beam therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **95** 488–497.

COOK, R. J. and LAWLESS, J. F. (2002). Analysis of repeated events. *Stat. Methods Med. Res.* **11** 141–166.

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

Key words and phrases. Accelerated failure time, Bayesian nonparametrics, chemoradiation, Dirichlet process, esophageal cancer, joint model, nonhomogeneous point process.

DE GRUTTOLA, V. and TU, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50** 1003–1014.

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

FAUCETT, C. L. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat. Med.* **15** 1663–1685.

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

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. and MOLENBERGHS, G., eds. (2009). *Longitudinal Data Analysis*. CRC Press, Boca Raton, FL. MR1500110

GEISSER, S. (1993). *Predictive Inference*: *An Introduction*. *Monographs on Statistics and Applied Probability* **55**. Chapman & Hall, New York. MR1252174

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. MR1278223

GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA.

GHOSH, S. K. and GHOSAL, S. (2006). Semiparametric accelerated failure time models for censored data. *Bayesian Stat. Appl.* **15** 213–229.

GHOSH, D. and LIN, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics* **56** 554–562. MR1795021

GRANDELL, J. (1976). *Doubly Stochastic Poisson Processes*. *Lecture Notes in Mathematics* **529**. Springer, Berlin. MR0433591

GUIDA, M., CALABRIA, R. and PULCINI, G. (1989). Bayes inference for a non-homogeneous Poisson process with power intensity law [reliability]. *IEEE Trans. Reliab.* **38** 603–609.

HATFIELD, L. A., BOYE, M. E. and CARLIN, B. P. (2011). Joint modeling of multiple longitudinal patient-reported outcomes and survival. *J. Biopharm. Statist.* **21** 971–991. MR2823361

HATFIELD, L. A., HODGES, J. S. and CARLIN, B. P. (2014). Joint models: When are treatment estimates improved? *Stat. Interface* **7** 439–453. MR3302373

HE, L., CHAPPLE, A., LIAO, Z., KOMAKI, R., THALL, P. F. and LIN, S. H. (2016). Bayesian regression analyses of radiation modality effects on pericardial and pleural effusion and survival in esophageal cancer. *Radiother. Oncol.* **121** 70–74.

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.

HUANG, C.-Y. and WANG, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *J. Amer. Statist. Assoc.* **99** 1153–1165. MR2109503

JARA, A., HANSON, T., QUINTANA, F., MUELLER, P. and ROSNER, G. (2018). Package 'DPpackage.'

JI, W., ZHENG, W., LI, B., CAO, C. and MAO, W. (2016). Influence of body mass index on the long-term outcomes of patients with esophageal squamous cell carcinoma who underwent esophagectomy as a primary treatment: A 10-year medical experience. *Medicine* **95** e4204.

KALBFLEISCH, J. D., SCHAUBEL, D. E., YE, Y. and GONG, Q. (2013). An estimating function approach to the analysis of recurrent and terminal events. *Biometrics* **69** 366–374. MR3071055

KUO, L. and YANG, T. Y. (1996). Bayesian computation for nonhomogeneous Poisson processes in software reliability. *J. Amer. Statist. Assoc.* **91** 763–773. MR1395743

LAWLESS, J. F. (1987). Regression methods for Poisson process data. *J. Amer. Statist. Assoc.* **82** 808–815. MR0909986

LEE, J., THALL, P. F. and LIN, S. H. (2019). Supplement to "Bayesian semiparametric joint regression analysis of recurrent adverse events and survival in esophageal cancer patients." DOI:10.1214/18-AOAS1182SUPP.

LEE, K. H., HANEUSE, S., SCHRAG, D. and DOMINICI, F. (2015). Bayesian semiparametric analysis of semicompeting risks data: Investigating hospital readmission after a pancreatic cancer diagnosis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 253–273. MR3302299

LEE, K. H., DOMINICI, F., SCHRAG, D. and HANEUSE, S. (2016). Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. *J. Amer. Statist. Assoc.* **111** 1075–1095. MR3561930

LI, Y., MÜLLER, P. and LIN, X. (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statist. Sinica* **21** 1201–1223. MR2827521

LIN, S. H., WANG, L., MYLES, B., THALL, P. F., HOFSTETTER, W. L., SWISHER, S. G., AJANI, J. A., COX, J. D., KOMAKI, R. and LIAO, Z. (2012). Propensity score-based comparison of long-term outcomes with 3-dimensional conformal radiotherapy vs intensity-modulated radiotherapy for esophageal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **84** 1078–1085.

LIU, L. and HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 65–81. MR2662234

LIU, L., WOLFE, R. A. and HUANG, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60** 747–756. MR2089451

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

MILLAR, R. B. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics* **65** 962–969. MR2649870

MÜLLER, P. and RODRIGUEZ, A. (2013). *Nonparametric Bayesian Inference. NSF-CBMS Regional Conference Series in Probability and Statistics* **9**. IMS, Beachwood, OH. MR3113683

OUYANG, B., SINHA, D., SLATE, E. H. and VAN BAKEL, A. B. (2013). Bayesian analysis of recurrent event with dependent termination: An application to a heart transplant study. *Stat. Med.* **32** 2629–2642. MR3067412

ROBERT, C. P. (2007). *The Bayesian Choice*: *From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. Springer, New York. MR2723361

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

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

SINHA, D., MAITI, T., IBRAHIM, J. G. and OUYANG, B. (2008). Current methods for recurrent events data with dependent termination: A Bayesian perspective. *J. Amer. Statist. Assoc.* **103** 866–878. MR2435473

SONG, X., DAVIDIAN, M. and TSIATIS, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58** 742–753. MR1945011

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

TORRE, L. A., BRAY, F., SIEGEL, R. L., FERLAY, J., LORTET-TIEULENT, J. and JEMAL, A. (2015). Global cancer statistics, 2012. *CA Cancer J. Clin.* **65** 87–108.

WALKER, S. and MALLICK, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics* **55** 477–483. MR1705102

WEN, S., HUANG, X., FRANKOWSKI, R. F., CORMIER, J. N. and PISTERS, P. (2016). A Bayesian multivariate joint frailty model for disease recurrences and survival. *Stat. Med.* **35** 4794–4812. MR3554994

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XU, Y., MÜLLER, P., WAHED, A. S. and THALL, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J. Amer. Statist. Assoc.* **111** 921–950. MR3561917

XU, G., CHIOU, S. H., HUANG, C.-Y., WANG, M.-C. and YAN, J. (2017). Joint scale-change models for recurrent events and failure time. *J. Amer. Statist. Assoc.* **112** 794–805. MR3671771

YE, W., LIN, X. and TAYLOR, J. M. G. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data—a two-stage regression calibration approach. *Biometrics* **64** 1238–1246. MR2522273

ZHANG, S. S., YANG, H., LUO, K. J., HUANG, Q. Y., CHEN, J. Y., YANG, F., CAI, X. L., XIE, X., LIU, Q. W., BELLA, A. E. et al. (2013). The impact of body mass index on complication and survival in resected oesophageal cancer: A clinical-based cohort and meta-analysis. *Br. J. Cancer* **109** 2894–2903.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 248–270
https://doi.org/10.1214/18-AOAS1186

A PENALIZED REGRESSION MODEL FOR THE JOINT ESTIMATION OF EQTL ASSOCIATIONS AND GENE NETWORK STRUCTURE[1]

BY MICOL MARCHETTI-BOWICK*,[2], YAOLIANG YU†, WEI WU* AND ERIC P. XING*

*Carnegie Mellon University** *and University of Waterloo*†

In this work, we present a new approach for jointly performing eQTL mapping and gene network inference while encouraging a transfer of information between the two tasks. We address this problem by formulating it as a multiple-output regression task in which we aim to learn the regression coefficients while simultaneously estimating the conditional independence relationships among the set of response variables. The approach we develop uses structured sparsity penalties to encourage the sharing of information between the regression coefficients and the output network in a mutually beneficial way. Our model, *inverse-covariance-fused lasso*, is formulated as a biconvex optimization problem that we solve via alternating minimization. We derive new, efficient optimization routines to solve each convex sub-problem that are based on extensions of state-of-the-art methods. Experiments on both simulated data and a yeast eQTL dataset demonstrate that our approach outperforms a large number of existing methods on the recovery of the true sparse structure of both the eQTL associations and the gene network. We also apply our method to a human Alzheimer's disease dataset and highlight some results that support previous discoveries about the disease.

REFERENCES

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

BANERJEE, S., YANDELL, B. S. and YI, N. (2008). Bayesian quantitative trait loci mapping for multiple traits. *Genetics* **179** 2275–2289.

BARABASI, A.-L. and OLTVAI, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5** 101–113.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

BREM, R. B. and KRUGLYAK, L. (2005). The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102** 1572–1577.

BREUNIG, J. S., HACKETT, S. R., RABINOWITZ, J. D. and KRUGLYAK, L. (2014). Genetic basis of metabolome variation in yeast. *PLoS Genet.* **10** e1004142.

Key words and phrases. eQTL mapping, gene network estimation, structured sparsity, multiple-output regression, covariance selection.

CHEN, X., KIM, S., LIN, Q., CARBONELL, J. G. and XING, E. P. (2010). Graph-structured multi-task regression and an efficient optimization method for general fused lasso. ArXiv preprint. Available at arXiv:1005.3579.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

FLUTRE, T., WEN, X., PRITCHARD, J. and STEPHENS, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9** e1003486.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GARDNER, T. S. and FAITH, J. J. (2005). Reverse-engineering transcription control networks. *Phys. Life Rev.* **2** 65–88.

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28** 27–30.

KIM, S., SOHN, K.-A. and XING, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* **25** i204–i212.

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

LEE, W. and LIU, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate Anal.* **111** 241–255. MR2944419

MALIK, M., CHILES III, J., XI, H. S., MEDWAY, C., SIMPSON, J., POTLURI, S., HOWARD, D., LIANG, Y., PAUMI, C. M., MUKHERJEE, S. et al. (2015). Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. *Hum. Mol. Genet.* **24** 3557–3570.

MARCHETTI-BOWICK, M., YU, Y., WU, W. and XING, E. P. (2019). Supplement to "A penalized regression model for the joint estimation of eQTL associations and gene network structure." DOI:10.1214/18-AOAS1186SUPP.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MICHAELSON, J. J., ALBERTS, R., SCHUGHART, K. and BEYER, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics* **11** 502.

NICA, A. C., MONTGOMERY, S. B., DIMAS, A. S., STRANGER, B. E., BEAZLEY, C., BARROSO, I. and DERMITZAKIS, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6** e1000895.

PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. MR2541591

RAI, P., KUMAR, A. and DAUME, H. (2012). Simultaneously leveraging output and task structures for multiple-output regression. In *Advances in Neural Information Processing Systems* (*NIPS*) 3185–3193.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

ROBERTS, C. J. and SELKER, E. U. (1995). Mutations affecting the biosynthesis of S-adenosylmethionine cause reduction of DNA methylation in Neurospora crassa. *Nucleic Acids Res.* **23** 4818–4826.

ROCKMAN, M. V. and KRUGLYAK, L. (2006). Genetics of global gene expression. *Nat. Rev. Genet.* **7** 862–872.

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19** 947–962. MR2791263

SOHN, K.-A. and KIM, S. (2012). Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 1081–1089.

STEPHENS, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8** e65245.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

TAN, K. M., LONDON, P., MOHAN, K., LEE, S.-I., FAZEL, M. and WITTEN, D. (2014). Learning graphical models with hubs. *J. Mach. Learn. Res.* **15** 3297–3331. MR3277170

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

WYTOCK, M. and KOLTER, Z. (2013). Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *International Conference on Machine Learning* (*ICML*) 1265–1273.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

YUAN, X.-T. and ZHANG, T. (2014). Partial Gaussian graphical model estimation. *IEEE Trans. Inform. Theory* **60** 1673–1687. MR3168429

ZHANG, Y. and YEUNG, D.-Y. (2010). A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* (*UAI'*10) 733–742. AUAI Press, Arlington, VA.

ZHANG, B., GAITERI, C., BODEA, L.-G., WANG, Z., MCELWEE, J., PODTELEZHNIKOV, A. A., ZHANG, C., XIE, T., TRAN, L., DOBRIN, R. et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153** 707–720.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 271–296
https://doi.org/10.1214/18-AOAS1192

A BAYESIAN RACE MODEL FOR RESPONSE TIMES UNDER CYCLIC STIMULUS DISCRIMINABILITY[1]

BY DEBORAH KUNKEL[*], KEVIN POTTER[†], PETER F. CRAIGMILE[*,2], MARIO PERUGGIA[*] AND TRISHA VAN ZANDT[*]

Ohio State University[*] and University of Massachusetts[†]

Response time (RT) data from psychology experiments are often used to validate theories of how the brain processes information and how long it takes a person to make a decision. When an RT results from a task involving two or more possible responses, the cognitive process that determines the RT may be modeled as the first-passage time of underlying competing (racing) processes with each process describing accumulation of information in favor of one of the responses. In one popular model the racers are assumed to be Gaussian diffusions. Their first-passage times are inverse Gaussian random variables and the resulting RT has a min-inverse Gaussian distribution. The RT data analyzed in this paper were collected in an experiment requiring people to perform a two-choice task in response to a regularly repeating sequence of stimuli. Starting from a min-inverse Gaussian likelihood for the RTs we build a Bayesian hierarchy for the rates and thresholds of the racing diffusions. The analysis allows us to characterize patterns in a person's sequence of responses on the basis of features of the person's diffusion rates (the "footprint" of the stimuli) and a person's gradual changes in speed as trends in the diffusion thresholds. Last, we propose that a small fraction of RTs arise from distinct, noncognitive processes that are included as components of a mixture model. In the absence of sharp prior information, the inclusion of these mixture components is accomplished via a two-stage, empirical Bayes approach. The resulting framework may be generalized readily to RTs collected under a variety of experimental designs.

REFERENCES

BAAYEN, R. H. and MILIN, P. (2010). Analyzing reaction times. *International Journal of Psychological Research* **3** 12–28.

BROWN, S. D. and HEATHCOTE, A. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation **57** 153–178.

CAPLIN, A. and MARTIN, D. (2016). The dual-process drift diffusion model: Evidence from response times. *Economic Inquiry* **54** 1274–1282.

CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika* **75** 613–632. MR2741490

HEITZ, R. P. and SCHALL, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron* **76** 616–628.

KIM, S., POTTER, K., CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2017). A Bayesian race model for recognition memory. *J. Amer. Statist. Assoc.* **112** 77–91. MR3646554

Key words and phrases. Cognitive modeling, inverse Gaussian distribution, Gaussian diffusion, harmonic regression, predictive diagnostics.

KUNKEL, D., POTTER, K., CRAIGMILE, P. F., PERUGGIA, M. and VAN ZANDT, T. (2019). Supplement to "A Bayesian race model for response times under cyclic stimulus discriminability." DOI:10.1214/18-AOAS1192SUPP.

LOGAN, G. D., VAN ZANDT, T., VERBRUGGEN, F. and WAGENMAKERS, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review* **121** 66–95.

LUCE, R. D. (1986). *Response Times*: *Their Role in Inferring Elementary Mental Organization*. Oxford Univ. Press, Oxford, UK.

NELSON, M. J., MURTHY, A. and SCHALL, J. D. (2016). Neural control of visual search by frontal eye field: Chronometry of neural events and race model processes. *J. Neurophysiol.* **115** 1954–1969.

RATCLIFF, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* **114** 510–532.

RATCLIFF, R. and MCKOON, G. (2007). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20** 873–922.

RATCLIFF, R., SMITH, P. L. and MCKOON, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Curr. Dir. Psychol. Sci.* **24** 458–470.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

USHER, M. and MCCLELLAND, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review* **108** 550–592.

VANDEKERCKHOVE, J. and TUERLINCKX, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychon. Bull. Rev.* **14** 1011–1026.

VAN ZANDT, T., COLONIUS, H. and PROCTOR, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychon. Bull. Rev.* **7** 208–256.

WAGENMAKERS, E.-J., FARRELL, S. and RATCLIFF, R. (2004). Estimation and interpretation of $1/f^{\alpha}$ noise in human cognition. *Psychon. Bull. Rev.* **11** 579–615.

WHELAN, R. (2008). Effective analysis of reaction time data. *Psychological Record* **58** 475–482.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 297–320
https://doi.org/10.1214/18-AOAS1199

BAYESIAN ANALYSIS OF INFANT'S GROWTH DYNAMICS WITH *IN UTERO* EXPOSURE TO ENVIRONMENTAL TOXICANTS[1]

BY JONGGYU BAEK, BIN ZHU AND PETER X. K. SONG

University of Michigan, University of Massachusetts Medical School and National Institutes of Health

Early infancy from at-birth to 3 years is critical for cognitive, emotional and social development of infants. During this period, infant's developmental tempo and outcomes are potentially impacted by *in utero* exposure to endocrine disrupting compounds (EDCs), such as bisphenol A (BPA) and phthalates. We investigate effects of ten ubiquitous EDCs on the infant growth dynamics of body mass index (BMI) in a birth cohort study. Modeling growth acceleration is proposed to understand the "force of growth" through a class of semiparametric stochastic velocity models. The great flexibility of such a dynamic model enables us to capture subject-specific dynamics of growth trajectories and to assess effects of the EDCs on potential delay of growth. We adopted a Bayesian method with the Ornstein–Uhlenbeck process as the prior for the growth rate function, in which the World Health Organization global infant's growth curves were integrated into our analysis. We found that BPA and most of phthalates exposed during the first trimester of pregnancy were inversely associated with BMI growth acceleration, resulting in a delayed achievement of infant BMI peak. Such early growth deficiency has been reported as a profound impact on health outcomes in puberty (e.g., timing of sexual maturation) and adulthood.

REFERENCES

AFEICHE, M., PETERSON, K. E., SANCHEZ, B. N., CANTONWINE, D., LAMADRID-FIGUEROA, H., SCHNAAS, L., ETTINGER, A. S., HERNANDEZ-AVILA, M., HU, H. and TELLEZ-ROJO, M. M. (2011). Prenatal lead exposure and weight of 0-to 5-year-old children in Mexico city. *Environ. Health Perspect.* **119** 1436–1441.

AGARWAL, D. K. and GELFAND, A. E. (2005). Slice sampling for simulation based fitting of spatial data models. *Stat. Comput.* **15** 61–69. MR2137218

BAEK, J., ZHU, B. and SONG, P. X. (2019). Supplement to "Bayesian analysis of infant's growth dynamics with *in utero* exposure to environmental toxicants." DOI:10.1214/18-AOAS1199SUPPA, DOI:10.1214/18-AOAS1199SUPPB.

BANISTER, C. E., KOESTLER, D. C., MACCANI, M. A., PADBURY, J. F., HOUSEMAN, E. A. and MARSIT, C. J. (2011). Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. *Epigenetics* **6** 920–927.

BINKIN, N. J., YIP, R., FLESHOOD, L. and TROWBRIDGE, F. L. (1988). Birth weight and childhood growth. *Pediatrics* **82** 828–834.

BOTTON, J., HEUDE, B., MACCARIO, J., DUCIMETIÈRE, P., CHARLES, M. A., BASDEVANT, A., BORYS, J. M., BRESSON, J. L., FROGUEL, P., LOMMEZ, A., OPPERT, J. M. and ROMON, M.

Key words and phrases. Body mass index, Markov chain Monte Carlo (MCMC), Ornstein–Uhlenbeck process, prenatal exposure, semiparametric stochastic velocity model.

(2008). Postnatal weight and height growth velocities at different ages between birth and 5 y and body composition in adolescent boys and girls. *Am. J. Clin. Nutr.* **87** 1760–1768.

BRAUN, J. M., JUST, A. C., WILLIAMS, P. L., SMITH, K. W., CALAFAT, A. M. and HAUSER, R. (2014). Personal care product use and urinary phthalate metabolite and paraben concentrations during pregnancy among women from a fertility clinic. *Journal of Exposure Science and Environmental Epidemiology* **24** 459–466.

CASALS-CASAS, C., FEIGE, J. N. and DESVERGNE, B. (2008). Interference of pollutants with PPARs: Endocrine disruption meets metabolism. *Int. J. Obes.* **32** S53–61.

COLE, T. J., DONALDSON, M. D. C. and BEN-SHLOMO, Y. (2010). SITAR-a useful instrument for growth curve analysis. *Int. J. Epidemiol.* **39** 1558–1566.

DIAMANTI-KANDARAKIS, E., BOURGUIGNON, J.-P., GIUDICE, L. C., HAUSER, R., PRINS, G. S., SOTO, A. M., ZOELLER, R. T. and GORE, A. C. (2009). Endocrine-disrupting chemicals: An endocrine society scientific statement. *Endocr. Rev.* **30** 293–342.

DURBÁN, M., HAREZLAK, J., WAND, M. P. and CARROLL, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Stat. Med.* **24** 1153–1167. MR2134571

GONZÁLEZ-COSSÍO, T., PETERSON, K. E., SANÍN, L. H., FISHBEIN, E., PALAZUELOS, E., ARO, A., HERNÁNDEZ-AVILA, M. and HU, H. (1997). Decrease in birth weight in relation to maternal bone-lead burden. *Pediatrics* **100** 856–862.

WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP (2006). *WHO Child Growth Standards*: *Length/Height-for-Age*, *Weight-for-Age*, *Weight-for-Length*, *Weight-for-Height and Body Mass Index-for-Age*: *Methods and Development*. World Health Organization, Geneva.

JENNS, R. M. and BAYLEY, N. (1937). A mathematical method for studying growth in children. *Hum. Neurobiol.* **9** 553–556.

JENSEN, S. M., RITZ, C., EJLERSKOV, K. T., MØLGAARD, C. and MICHAELSEN, K. F. (2015). Infant BMI peak, breastfeeding, and body composition at age 3 y. *Am. J. Clin. Nutr.* **101** 319–325.

JONES-SMITH, J. C., NEUFELD, L. M., LARAIA, B., RAMAKRISHNAN, U., GARCIA-GUERRA, A. and FERNALD, L. C. H. (2013). Early life growth trajectories and future risk for overweight. *Nutr. Diabetes* **3** e60.

KOBROSLY, R. W., PARLETT, L. E., STAHLHUT, R. W., BARRETT, E. S. and SWAN, S. H. (2012). Socioeconomic factors and phthalate metabolite concentrations among United States women of reproductive age. *Environ. Res.* **115** 11–17.

LÓPEZ-PINTADO, S. and MCKEAGUE, I. W. (2013). Recovering gradients from sparsely observed functional data. *Biometrics* **69** 396–404. MR3071058

MARIE, C., VENDITTELLI, F. and SAUVANT-ROCHAT, M. P. (2015). Obstetrical outcomes and biomarkers to assess exposure to phthalates: A review. *Environ. Int.* **83** 116–136.

MARSEE, K., WOODRUFF, T. J., AXELRAD, D. A., CALAFAT, A. M. and SWAN, S. H. (2006). Estimated daily phthalate exposures in a population of mothers of male infants exhibiting reduced anogenital distance. *Environ. Health Perspect.* **114** 805–809.

MCKEAGUE, I. W., LOPEZ-PINTADO, S., HALLIN, M. and SIMAN, M. (2011). Analyzing growth trajectories. *Journal of Developmental Origins of Health and Disease* **2** 322–329.

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. MR1994729

NHANES, I. (2009). Fourth national report on human exposure to environmental chemicals. In *Department of Health and Human Services Centers for Disease Control and Prevention*, Atlanta, GA.

PIAGET, J. (2000). Piaget's theory of cognitive development. In *Childhood Cognitive Development*: *The Essential Readings* 33–47.

PREECE, M. A. and BAINES, M. J. (1978). A new family of mathematical models describing the human growth curve. *Ann. Hum. Biol.* **5** 1–24.

RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314

SAARNI, C. (1999). *The Development of Emotional Competence*. Guilford, New York.
SCHETTLER, T. (2006). Human exposure to phthalates via consumer products. *Int. J. Androl.* **29** 134–139.
SILVERWOOD, R., DE STAVOLA, B., COLE, T. and LEON, D. (2005). BMI peak in infancy as a predictor for later BMI in the uppsala family study. *Int. J. Obes.* **33** 929–937.
TAYLOR, R. W., GRANT, A. M., GOULDING, A. and WILLIAMS, S. M. (2005). Early adiposity rebound: Review of papers linking this to subsequent obesity in children and adults. *Curr. Opin. Clin. Nutr. Metab. Care* **8** 607–612.
TÉLLEZ-ROJO, M. M., HERNÁNDEZ-AVILA, M., LAMADRID-FIGUEROA, H., SMITH, D., HERNÁNDEZ-CADENA, L., MERCADO, A., ARO, A., SCHWARTZ, J. and HU, H. (2004). Impact of bone lead and bone resorption on plasma and whole blood lead levels during pregnancy. *Am. J. Epidemiol.* **160** 668–678.
WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372. MR0522220
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
ZHANG, A., HU, H., SÁNCHEZ, B. N., ETTINGER, A. S., PARK, S. K., CANTONWINE, D., SCHNAAS, L., WRIGHT, R. O., LAMADRID-FIGUEROA, H. and TELLEZ-ROJO, M. M. (2012). Association between prenatal lead exposure and blood pressure in children. *Environ. Health Perspect.* **120**.
ZHAO, Y., SHI, H. J., XIE, C. M., CHEN, J., LAUE, H. and ZHANG, Y. H. (2015). Prenatal phthalate exposure, infant growth, and global DNA methylation of human placenta. *Environ. Mol. Mutagen.* **56** 286–292.
ZHU, B., TAYLOR, J. M. G. and SONG, P. X.-K. (2011). Semiparametric stochastic modeling of the rate function in longitudinal studies. *J. Amer. Statist. Assoc.* **106** 1485–1495. MR2896851

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 321–339
https://doi.org/10.1214/18-AOAS1187

JOINT MEAN AND COVARIANCE MODELING OF MULTIPLE HEALTH OUTCOME MEASURES

BY XIAOYUE NIU[1] AND PETER D. HOFF[2]

Pennsylvania State University and Duke University

Health exams determine a patient's health status by comparing the patient's measurement with a population reference range, a 95% interval derived from a homogeneous reference population. Similarly, most of the established relation among health problems are assumed to hold for the entire population. We use data from the 2009–2010 National Health and Nutrition Examination Survey (NHANES) on four major health problems in the U.S. and apply a joint mean and covariance model to study how the reference ranges and associations of those health outcomes could vary among subpopulations. We discuss guidelines for model selection and evaluation, using standard criteria such as AIC in conjunction with posterior predictive checks. The results from the proposed model can help identify subpopulations in which more data need to be collected to refine the reference range and to study the specific associations among those health problems.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csaki, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

BOIK, R. J. (2002). Spectral models for covariance matrices. *Biometrika* **89** 159–182. MR1888370

BOIK, R. J. (2003). Principal component models for correlation matrices. *Biometrika* **90** 679–701. MR2006844

CDC/NCHS (2010a). National Health and Nutrition Examination Survey Data, 2009–2010. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

CDC/NCHS (2010b). National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD.

CHIU, T. Y. M., LEONARD, T. and TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074

CLSI (2008). *Defining*, *Establishing*, *and Verifying Reference Intervals in the Clinical Laboratory*: *Approved Guideline*, 3rd ed. CLSI document EP28-A3c. Clinical and Laboratory Standards Institute, Wayne, PA.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39 (with a discussion). MR0893334

CRIPPS, E., CARTER, C. and KOHN, R. (2005). Variable selection and covariance selection in multivariate regression models. In *Bayesian Thinking*: *Modeling and Computation* (D. Dey and C. R. Rao, eds.). *Handbook of Statist.* **25** 519–552. Elsevier/North-Holland, Amsterdam. MR2490538

Key words and phrases. Heterogeneous population, reference range, covariance regression, NHANES.

ENGLE, R. F. and KRONER, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory* **11** 122–150. MR1325104

FONG, P. W., LI, W. K. and AN, H.-Z. (2006). A simple multivariate ARCH model specified by random coefficients. *Comput. Statist. Data Anal.* **51** 1779–1802. MR2307543

FOULDS, H., BREDIN, S. and WARBURTON, D. (2012). The relationship between diabetes and obesity across different ethnicities. *J. Diabetes Metab.* **3**.

FRASER, S. D. S., RODERICK, P. J., MCLNTYRE, N. J., HARRIS, S., MCLNTYRE, C. W., FLUCK, R. J. and TAAL, M. W. (2012). Socio-economic disparities in the distribution of cardiovascular risk in chronic kidney disease stage 3. *Nephron, Clin. Pract.* **122** 58–65.

GASKINS, J. T. and DANIELS, M. J. (2013). A nonparametric prior for simultaneous covariance estimation. *Biometrika* **100** 125–138. MR3034328

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

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. Ser. B* **29** 83–100. MR0216699

HARRIS, E. K. and BOYD, J. C. (1990). On dividing reference data into subgroups to produce separate reference ranges. *Clin. Chem.* **36** 265–270.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 971–992. MR2750253

HOFF, P. D. and NIU, X. (2012). A covariance regression model. *Statist. Sinica* **22** 729–753. MR2954359

KDIGO (2013). Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Inter., Suppl.* **3** 1–150.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

MATTIX, H. J., HSU, C.-Y., SHAYKEVICH, S. and CURHAN, G. (2002). Use of the albumin/creatinine ratio to detect microalbuminuria: Implications of sex and race. *J. Am. Soc. Nephrol.* **13** 1034–1039.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. [Second edition of MR0727836.] MR3223057

NIDDK (2013). U.S. Renal Data System, USRDS 2013 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

NIU, X. and HOFF, P. D. (2019). Supplement to "Joint mean and covariance modeling of multiple health outcome measures." DOI:10.1214/18-AOAS1187SUPP.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

WINSHIP, C. and RADBILL, L. (1994). Sampling weights and regression analysis. *Sociol. Methods Res.* **23** 230–257.

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 340–366
https://doi.org/10.1214/18-AOAS1188

BAYESIAN LATENT HIERARCHICAL MODEL FOR TRANSCRIPTOMIC META-ANALYSIS TO DETECT BIOMARKERS WITH CLUSTERED META-PATTERNS OF DIFFERENTIAL EXPRESSION SIGNALS

BY ZHIGUANG HUO[1], CHI SONG[2] AND GEORGE TSENG[1,2]

University of Florida, Ohio State University and University of Pittsburgh

Due to the rapid development of high-throughput experimental techniques and fast-dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase the statistical power to detect disease-related biomarkers. In this paper we introduce a Bayesian latent hierarchical model to perform transcriptomic meta-analysis. This method is capable of detecting genes that are differentially expressed (DE) in only a subset of the combined studies, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and three examples, including a microarray dataset from metabolism-related knockout mice, an RNA-seq dataset from HIV transgenic rats and cross-platform datasets from human breast cancer are used to demonstrate the performance of the proposed method.

REFERENCES

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** Art. ID R106.

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERGER, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*, Springer, New York. MR0804611

BHATTACHARJEE, S., RAJARAMAN, P., JACOBS, K. B., WHEELER, W. A., MELIN, B. S., HARTGE, P., YEAGER, M., CHUNG, C. C., CHANOCK, S. J., CHATTERJEE, N. et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet*. **90** 821–835.

BIRNBAUM, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc*. **49** 559–574. MR0065101

CHANG, L.-C., LIN, H.-M., SIBILLE, E. and TSENG, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: Comparisons, statistical characterization and an application guideline. *BMC Bioinform*. **14** Art. ID 368.

Key words and phrases. Transcriptomic differential analysis, meta-analysis, Bayesian hierarchical model, Dirichlet process.

COOPER, H., HEDGES, L. V. and VALENTINE, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, New York.

DOMANY, E. (2014). Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. *Cancer Res*. **74** 4612–4621.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc*. **99** 96–104. MR2054289

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci*. **23** 1–22. MR2431866

EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc*. **104** 1015–1028. MR2562003

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol*. **23** 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc*. **96** 1151–1160. MR1946571

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

FISHER, R. A. (1934). *Statistical Methods for Research Workers*. Hafner Publishing Co., New York. MR0346954

GENTLEMAN, R., CAREY, V. J., HUBER, W., IRIZARRY, R. A. and DUDOIT, S., eds. (2006). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York. MR2201836

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist*. **27** 143–158. MR1701105

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

HUO, Z., SONG, C. and TSENG, G. (2019). Supplement to "Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals." DOI:10.1214/18-AOAS1188SUPPA, DOI:10.1214/18-AOAS1188SUPPB, DOI:10.1214/18-AOAS1188SUPPC.

JACOB, L., GAGNON-BARTSCH, J. A. and SPEED, T. P. (2016). Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* **17** 16–28. MR3449847

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

KANG, H. M., YE, C. and ESKIN, E. (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180** 1909–1925.

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3** Art. ID e161.

LI, J. and TSENG, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat*. **5** 994–1019. MR2840184

LI, M. D., CAO, J., WANG, S., WANG, J., SARKAR, S., VIGORITO, M., MA, J. Z. and CHANG, S. L. (2013). Transcriptome sequencing of gene expression in the brain of the HIV-1 transgenic rat. *PLoS ONE* **8** Art. ID e59582.

LI, Q., WANG, S., HUANG, C.-C., YU, M. and SHAO, J. (2014). Meta-analysis based variable selection for gene expression data. *Biometrics* **70** 872–880. MR3295748

LISTGARTEN, J., KADIE, C., SCHADT, E. E. and HECKERMAN, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. USA* **107** 16465–16470.

LITTELL, R. C. and FOLKS, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Amer. Statist. Assoc.* **66** 802–806. MR0312634

MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110. MR2082149

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. MR2758178

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

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

OHIO SUPERCOMPUTER CENTER (1987). Ohio Supercomputer Center. Available at http://osc.edu/ark:/19495/f5s1ph73.

QUINLAN, A. R. and HALL, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26** 841–842.

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** Art. ID e184.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SCHARPF, R. B., TJELMELAND, H., PARMIGIANI, G. and NOBEL, A. B. (2009). A Bayesian model for cross-study differential gene expression. *J. Amer. Statist. Assoc.* **104** 1295–1310. MR2750563

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.

SMYTH, G. K. (2005). Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420. Springer, New York.

SONG, C. and TSENG, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* **8** 777–800. MR3262534

STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS JR., R. M. (1949). *The American Soldier*: *Adjustment During Army Life*. Princeton Univ. Press, Princeton, NJ.

TRAPNELL, C., PACHTER, L. and SALZBERG, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25** 1105–1111.

TSENG, G. C., 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

TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** 5116.

WALKER, W. L., LIAO, I. H., GILBERT, D. L., WONG, B., POLLARD, K. S., MCCULLOCH, C. E., LIT, L. and SHARP, F. R. (2008). Empirical Bayes accomodation of batch-effects in microarray data using identical replicate reference samples: Application to RNA expression profiling of blood from Duchenne muscular dystrophy patients. *BMC Genomics* **9** Art. ID 494.

WEISS, R. A. (1993). How does HIV cause AIDS? *Science* **260** 1273–1279.

ZHAO, Y., KANG, J. and YU, T. (2014). A Bayesian nonparametric mixture model for selecting genes and gene subnetworks. *Ann. Appl. Stat.* **8** 999–1021. MR3262543

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 367–392
https://doi.org/10.1214/18-AOAS1189

MODELING WITHIN-HOUSEHOLD ASSOCIATIONS IN HOUSEHOLD PANEL STUDIES

BY FIONA STEELE*, PAUL S. CLARKE†,[1] AND JOUNI KUHA*

London School of Economics & Political Science and University of Essex†*

Household panel data provide valuable information about the extent of similarity in coresidents' attitudes and behaviours. However, existing analysis approaches do not allow for the complex association structures that arise due to changes in household composition over time. We propose a flexible marginal modeling approach where the changing correlation structure between individuals is modeled directly and the parameters estimated using second-order generalized estimating equations (GEE2). A key component of our correlation model specification is the "superhousehold", a form of social network in which pairs of observations from different individuals are connected (directly or indirectly) by coresidence. These superhouseholds partition observations into clusters with nonstandard and highly variable correlation structures. We thus conduct a simulation study to evaluate the accuracy and stability of GEE2 for these models. Our approach is then applied in an analysis of individuals' attitudes towards gender roles using British Household Panel Survey data. We find strong evidence of between-individual correlation before, during and after coresidence, with large differences among spouses, parent–child, other family, and unrelated pairs. Our results suggest that these dependencies are due to a combination of nonrandom sorting and causal effects of coresidence.

REFERENCES

ATKINS, D. C. (2005). Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. *J. Fam. Psychol.* **19** 98–110.

BALLAS, D. and TRANMER, M. (2012). Happy people or happy places? A multilevel modeling approach to the analysis of happiness and well-being. *Int. Reg. Sci. Rev.* **35** 70–102.

BAUER, D. J., GOTTFREDSON, N. C., DEAN, D. and ZUCKER, R. A. (2013). Analyzing repeated measures data on individuals nested within groups: Accounting for dynamic group effects. *Psychol. Methods* **18** 1–14.

BERRIDGE, D., PENN, R. and GANJALI, M. (2009). Changing attitudes to gender roles: A longitudinal analysis of ordinal response data from the British household panel study. *Int. Sociol.* **24** 346–367.

BLACKWELL, D. L. and LICHTER, D. T. (2004). Homogamy among dating, cohabiting and married couples. *Sociol. Q.* **45** 719–737.

BRYNIN, M., LONGHI, S. and MARTÍNEZ PÉREZ, Á. (2008). The social significance of homogamy. 73–90 5. Routledge, New York.

BUCK, N. and MCFALL, S. (2012). Understanding society: Design overview. *Longitud. Life Course Stud.* **3** 5–17.

Key words and phrases. Household effects, household correlation, longitudinal households, homophily, multiple membership multilevel model, marginal model, generalised estimating equations.

BUTTERWORTH, P. and RODGERS, B. (2006). Concordance in the mental health of spouses: Analysis of a large national household panel survey. *Psychol. Med.* **36** 685–697.

CHAGANTY, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plann. Inference* **63** 39–54. MR1474184

CHANDOLA, T., BARTLEY, M., WIGGINS, R. and SCHOFIELD, P. (2003). Social inequalities in health by individual and household measures of social position in a cohort of healthy people. *J. Epidemiol. Community Health* **57** 56–62.

CHIU, T. Y. M., LEONARD, T. and TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074

CROWDER, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82** 407–410.

DAVILLAS, A. and PUDNEY, S. (2017). Concordance of health states in couples: Analysis of self-reported, nurse administered and blood-based biomarker data in the UK understanding society panel. *J. Health Econ.* **56** 87–102.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28** 157–175.

DUNCAN, G. and HILL, M. (1985). Conceptions of longitudinal households: Fertile or futile? *J. Econ. Soc. Meas.* **13** 361–375.

FOWLER, J. H. and CHRISTAKIS, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study. *Br. Med. J.* **337** a2338.

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

GOLDSTEIN, H. (2010). *Multilevel Statistical Models*, 4th ed. Wiley, London.

GOLDSTEIN, H., RASBASH, J., BROWNE, W. J., WOODHOUSE, G. and POULAIN, M. (2000). Multilevel models in the study of dynamic household structures. *Eur. J. Popul.* **16** 373–387.

HARDIN, J. W. and HILBE, J. M. (2013). *Generalized Estimating Equations*, 2nd ed. CRC Press, Boca Raton, FL. MR3134775

HØJSGAARD, S., HALEKOH, U. and YAN, J. (2006). The R package `geepack` for generalized estimating equations. *J. Stat. Softw.* **15** 1–11.

INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH (ISER) (2009). *British Household Panel Survey*: *Waves* 1–17, 1991–2008, 6th ed. Univ. Essex, Institute for Social and Economic Research [original data producer(s)], Colchester, Essex. UK Data Archive [distributor]. SN: 5151.

JENNRICH, R. I. and SCHLUCHTER, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805–820. MR0872961

JOHNSTON, R., JONES, K., PROPPER, C., SARKER, R., BURGESS, S. and BOLSTER, A. (2005). A missing level in the analyses of British voting behaviour: The household as context as shown by analyses of a 1992–1997 longitudinal survey. *Elect. Stud.* **24** 201–225.

JONES, B. and WEST, M. (2005). Covariance decomposition in undirected Gaussian graphical models. *Biometrika* **92** 779–786. MR2234185

KALMIJN, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. *Annu. Rev. Sociol.* **24** 395–421.

KEIZER, R. and SCHENK, N. (2012). Becoming a parent and relationship satisfaction: A longitudinal dyadic perspective. *J. Marriage Fam.* **74** 759–773.

KUK, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika* **94** 939–952. MR2416800

KUK, A. Y. C. and NOTT, D. J. (2000). A pairwise likelihood approach to analysing correlated binary data. *Statist. Probab. Lett.* **47** 329–335.

LECKIE, G. and GOLDSTEIN, H. (2009). The limitations of using school league tables to inform school choice. *J. Roy. Statist. Soc. Ser. A* **172** 835–851. MR2751830

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LIANG, K.-Y., ZEGER, S. L. and QAQISH, B. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B* **54** 3–40. MR1157713

MCPHERSON, M., SMITH-LOVIN, L. and COOK, J. M. (2001). Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27** 415–444.

MILNER, A., SPITTAL, M. J., PAGE, A. and LAMONTAGNE, A. D. (2014). The effect of leaving employment on mental health: Testing 'adaptation' versus 'sensitisation' in a cohort of working-age australians. *Occup. Environ. Med.* **71** 167–174.

MURPHY, M. J. (1996). The dynamic household as a logical concept and its use in demography. *Eur. J. Popul.* **12** 363–381.

PEARSON, M. and WEST, P. (2003). Drifting smoke rings: Social network analysis and Markov processes in a longitudinal study of friendship groups and risk-taking. *Connections* **25** 59–76.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

PRENTICE, R. L. and ZHAO, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47** 825–839. MR1141951

RAUDENBUSH, S. W., BRENNAN, R. T. and BARNETT, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *J. Fam. Psychol.* **9** 161–174.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

SACKER, A., WIGGINS, R. and BARTLEY, M. (2006). Time and place: Putting individual health into context. A multilevel analysis of the British household panel survey, 1991–2001. *Health Place* **12** 279–290.

SHULTS, J. and HILBE, J. M. (2014). *Quasi-Least Squares Regression. Monographs on Statistics and Applied Probability* **132**. Chapman and Hall/CRC Press, Boca Raton, FL. MR3202387

STEELE, F., CLARKE, P. S. and KUHA, J. (2019). Supplement to "Modeling within-household associations in household panel studies." DOI:10.1214/18-AOAS1189SUPP.

SWEETING, H., BHASKAR, A., BENZEVAL, M., POPHAM, F. and HUNT, K. (2014). Changing gender roles and attitudes and their implications for well-being around the new millennium. *Soc. Psychiatry Psychiatr. Epidemiol.* **49** 791–809.

YAN, J. and FINE, J. (2004). Estimating equations for association structures. *Stat. Med.* **23** 859–874.

ZIEGLER, A., KASTNER, C. and BLETTNER, M. (1998). The generalised estimating equations: An annotated bibliography. *Biom. J.* **40** 115–139. MR1625993

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 393–419
https://doi.org/10.1214/18-AOAS1195

FRÉCHET ESTIMATION OF TIME-VARYING COVARIANCE MATRICES FROM SPARSE DATA, WITH APPLICATION TO THE REGIONAL CO-EVOLUTION OF MYELINATION IN THE DEVELOPING BRAIN

BY ALEXANDER PETERSEN, SEAN DEONI[1] AND HANS-GEORG MÜLLER[2]

University of California, Santa Barbara, Brown University and University of California, Davis

Assessing brain development for small infants is important for determining how the human brain grows during the early period of life when the rate of brain growth is at its peak. The development of MRI techniques has enabled the quantification of brain development. A key quantity that can be extracted from MRI measurements is the level of myelination, where myelin acts as an insulator around nerve fibers and its deployment makes nerve pulse propagation more efficient. The co-variation of myelin deployment across different brain regions provides insights into the co-development of brain regions and can be assessed as correlation matrix that varies with age. Typically, available data for each child are very sparse, due to the cost and logistic difficulties of arranging MRI brain scans for infants. We showcase here a method where data per subject are limited to measurements taken at only one random age, so that one has cross-sectional data available, while aiming at the time-varying dynamics. This situation is encountered more generally in cross-sectional studies where one observes p-dimensional vectors at one random time point per subject and is interested in the $p \times p$ correlation matrix function over the time domain. The challenge is that at each observation time one observes only a p-vector of measurements but not a covariance or correlation matrix. For such very sparse data, we develop a Fréchet estimation method. Given a metric on the space of covariance matrices, the proposed method generates a matrix function where at each time the matrix is a nonnegative definite covariance matrix, for which we demonstrate consistency properties. We discuss how this approach can be applied to myelin data in the developing brain and what insights can be gained.

REFERENCES

ŞENTÜRK, D. and MÜLLER, H.-G. (2010). Functional varying coefficient models for longitudinal data. *J. Amer. Statist. Assoc.* **105** 1256–1264. MR2752619

ŞENTÜRK, D. and NGUYEN, D. V. (2011). Varying coefficient models for sparse noise-contaminated longitudinal data. *Statist. Sinica* **21** 1831–1856. MR2896001

ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes. Probability and Mathematical Statistics* **27**. Academic Press, New York. MR0448463

Key words and phrases. Myelination, neurocognitive development, MRI, covariance, correlation analysis, time dynamics, random objects, local smoothing.

BARTZOKIS, G., LU, P., TINGUS, K., MENDEZ, M., RICHARD, A., PETERS, D. G. et al. (2010). Lifespan trajectory of myelin integrity and maximum motor speed. *Neurobiol. Aging* **31** 1554–1562.

BILENBERG, N. (1999). The child behavior checklist (CBCL) and related material: Standardization and validation in Danish population based and clinically based samples. *Acta Psychiatr. Scand.* **100** 2–52.

BRODY, B. A., KINNEY, H. C., KLOMAN, A. S. and GILLES, F. H. (1987). Sequence of central nervous system myelination in human infancy. I. An autopsy study of myelination. *J. Neuropathol. Exp. Neurol.* **46** 283–301.

CARMICHAEL, O., CHEN, J., PAUL, D. and PENG, J. (2013). Diffusion tensor smoothing through weighted Karcher means. *Electron. J. Stat.* **7** 1913–1956. MR3084676

CASEY, B., GALVAN, A. and HARE, T. A. (2005). Changes in cerebral functional organization during cognitive development. *Curr. Opin. Neurobiol.* **15** 239–244.

CHIOU, J.-M. and MÜLLER, H.-G. (2016). A pairwise interaction model for multivariate functional and longitudinal data. *Biometrika* **103** 377–396. MR3509893

CHLEBOWSKI, C., ROBINS, D. L., BARTON, M. L. and FEIN, D. (2013). Large-scale use of the modified checklist for autism in low-risk toddlers. *Pediatrics* **131** e1121–e1127.

DAI, X., MÜLLER, H.-G., WANG, J.-L. and DEONI, S. C. (2017a). Age-dynamic networks and functional correlation for early white matter myelination. Preprint.

DAI, X., HADJIPANTELIS, P., WANG, J.-L., DEONI, S. C. and MÜLLER, H.-G. (2017b). Longitudinal associations between white matter maturation and cognitive development across early childhood. Preprint.

DEONI, S. C. (2007). High-resolution T1 mapping of the brain at 3T with driven equilibrium single pulse observation of T1 with high-speed incorporation of RF field inhomogeneities (DESPOT1-HIFI). *J. Magn. Reson. Imaging* **26** 1106–1111.

DEONI, S. C. (2011). Correction of main and transmit magnetic field (B0 and B1) inhomogeneity effects in multicomponent-driven equilibrium single-pulse observation of T1 and T2. *Magn. Reson. Med.* **65** 1021–1035.

DEONI, S. C. and KOLIND, S. H. (2015). Investigating the stability of mcDESPOT myelin water fraction values derived using a stochastic region contraction approach. *Magn. Reson. Med.* **73** 161–169.

DEONI, S. C., PETERS, T. M. and RUTT, B. K. (2004). Determination of optimal angles for variable nutation proton magnetic spin-lattice, T1, and spin-spin, T2, relaxation times measurement. *Magn. Reson. Med.* **51** 194–199.

DEONI, S. C., RUTT, B. K. and PETERS, T. M. (2003). Rapid combined T1 and T2 mapping using gradient recalled acquisition in the steady state. *Magn. Reson. Med.* **49** 515–526.

DEONI, S. C., RUTT, B. K. and PETERS, T. M. (2006). Synthetic T1-weighted brain image generation with incorporated coil intensity correction using DESPOT1. *Magn. Reson. Imaging* **24** 1241–1248.

DEONI, S. C., WARD, H. A., PETERS, T. M. and RUTT, B. K. (2004). Rapid T2 estimation with phase-cycled variable nutation steady-state free precession. *Magn. Reson. Med.* **52** 435–439.

DEONI, S. C., WILLIAMS, S. C., JEZZARD, P., SUCKLING, J., MURPHY, D. G. and JONES, D. K. (2008a). Standardized structural magnetic resonance imaging in multicentre studies using quantitative T1 and T2 imaging at 1.5 T. *NeuroImage* **40** 662–671.

DEONI, S. C., RUTT, B. K., ARUN, T., PIERPAOLI, C. and JONES, D. K. (2008b). Gleaning multicomponent T1 and T2 information from steady-state imaging data. *Magn. Reson. Med.* **60** 1372–1387.

DEONI, S. C., MERCURE, E., BLASI, A., GASSTON, D., THOMSON, A., JOHNSON, M., WILLIAMS, S. C. and MURPHY, D. G. (2011). Mapping infant brain myelination with magnetic resonance imaging. *J. Neurosci.* **31** 784–791.

DEONI, S. C., O'MUIRCHEARTAIGH, J., ELISON, J. T., WALKER, L., DOERNBERG, E., WASKIEWICZ, N., DIRKS, H., PIRYATINSKY, I., DEAN III, D. C. and JUMBE, N. (2016). White matter maturation profiles through early childhood predict general cognitive ability. *Brain Struct. Funct.* **221** 1189–1203.

DEONI, S. C., DEAN, D. C., REMER, J., DIRKS, H. and O'MUIRCHEARTAIGH, J. (2015). Cortical maturation and myelination in healthy toddlers and young children. *NeuroImage* **115** 147–161.

ERUS, G., BATTAPADY, H., SATTERTHWAITE, T. D., HAKONARSON, H., GUR, R. E., DAVATZIKOS, C. and GUR, R. C. (2014). Imaging patterns of brain development and their relationship to cognition. *Cereb. Cortex* **25** 1676–1684.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. Chapman & Hall, London. MR1383587

FIELDS, R. (2005). Myelination: An overlooked mechanism of synaptic plasticity? *Neuroscientist* **11** 528–531.

FIELDS, R. D. (2008). White matter in learning, cognition and psychiatric disorders. *Trends Neurosci.* **31** 361–370.

FLYNN, S., LANG, D., MACKAY, A., GOGHARI, V., VAVASOUR, I., WHITTALL, K., SMITH, G., ARANGO, V., MANN, J., DWORK, A. et al. (2003). Abnormalities of myelination in schizophrenia detected in vivo with MRI, and post-mortem with analysis of oligodendrocyte proteins. *Mol. Psychiatry* **8** 811–820.

FORNARI, E., KNYAZEVA, M. G., MEULI, R. and MAEDER, P. (2007). Myelination shapes functional activity in the developing brain. *NeuroImage* **38** 511–518.

FRYER, S. L., FRANK, L. R., SPADONI, A. D., THEILMANN, R. J., NAGEL, B. J., SCHWEINSBURG, A. D. and TAPERT, S. F. (2008). Microstructural integrity of the corpus callosum linked with neuropsychological performance in adolescents. *Brain Cogn.* **67** 225–233.

GRYDELAND, H., WALHOVD, K. B., TAMNES, C. K., WESTLYE, L. T. and FJELL, A. M. (2013). Intracortical myelin links with performance variability across the human lifespan: Results from T1- and T2-weighted MRI myelin mapping and diffusion tensor imaging. *J. Neurosci.* **33** 18618–18630.

HAGMANN, P., SPORNS, O., MADAN, N., CAMMOUN, L., PIENAAR, R., WEDEEN, V. J., MEULI, R., THIRAN, J.-P. and GRANT, P. (2010). White matter maturation reshapes structural connectivity in the late developing human brain. *Proc. Natl. Acad. Sci. USA* **107** 19067–19072.

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester. MR3379106

ISHIBASHI, T., DAKIN, K., STEVENS, B., LEE, P., KOZLOV, S. and STEWART, C. E. A. (2006). Astrocytes promote myelination in response to electrical impulses. *Neuron* **49** 823–832.

KNICKMEYER, R. C., GOUTTARD, S., KANG, C., EVANS, D., WILBER, K., SMITH, J. K., HAMER, R. M., LIN, W., GERIG, G. and GILMORE, J. H. (2008). A structural MRI study of human brain development from birth to 2 years. *J. Neurosci.* **28** 12176–12182.

LANG, D., YIP, E., MACKAY, A., THORNTON, A., VILA-RODRIGUEZ, F. and MACEWAN, G. E. A. (2003). Abnormalities of myelination in schizophrenia detected in vivo with MRI, and post-mortem with analysis of oligodendrocyte proteins. *Mol. Psychiatry* **8** 811–820.

LEBEL, C. and BEAULIEU, C. (2011). Longitudinal development of human brain wiring continues from childhood into adulthood. *J. Neurosci.* **31** 10937–10947.

LENROOT, R. K. and GIEDD, J. N. (2006). Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neurosci. Biobehav. Rev.* **30** 718–729.

MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415. MR0679685

MILLER, D. J., DUKA, T., STIMPSON, C. D., SCHAPIRO, S. J., BAZE, W. B., MCARTHUR, M. J., FOBBS, A. J., SOUSA, A. M., ŠESTAN, N., WILDMAN, D. E. et al. (2012). Prolonged myelination in human neocortical evolution. *Proc. Natl. Acad. Sci. USA* **109** 16480–16485.

MULLEN, E. M. et al. (1995). *Mullen Scales of Early Learning*. AGS, Circle Pines, MN.
NAGY, Z., WESTERBERG, H. and KLINGBERG, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *J. Cogn. Neurosci.* **16** 1227–1233.
PAUS, T., COLLINS, D., EVANS, A., LEONARD, G., PIKE, B. and ZIJDENBOS, A. (2001). Maturation of white matter in the human brain: A review of magnetic resonance studies. *Brain Res. Bull.* **54** 255–266.
PETERSEN, A., DEONI, S. and MÜLLER, H.-G. (2019). Supplement to "Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional co-evolution of myelination in the developing brain." DOI:10.1214/18-AOAS1195SUPP.
PETERSEN, A. and MÜLLER, H.-G. (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika* **103** 103–120. MR3465824
PETERSEN, A. and MÜLLER, H.-G. (2018). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* To appear. Available at arXiv:1608.03012.
PIGOLI, D., ASTON, J. A. D., DRYDEN, I. L. and SECCHI, P. (2014). Distances and inference for covariance operators. *Biometrika* **101** 409–422. MR3215356
RODIER, P. M. (1995). Developing brain as a target of toxicity. *Environ. Health Perspect.* **103** 73–76.
SHAFEE, R., BUCKNER, R. L. and FISCHL, B. (2015). Gray matter myelination of 1555 human brains using partial volume corrected MRI images. *NeuroImage* **105** 473–485.
SHAW, P., GREENSTEIN, D., LERCH, J., CLASEN, L., LENROOT, R., GOGTAY, N., EVANS, A., RAPOPORT, J. and GIEDD, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature* **440** 676–679.
SHAW, P., KABANI, N. J., LERCH, J. P., ECKSTRAND, K., LENROOT, R., GOGTAY, N., GREENSTEIN, D., CLASEN, L., EVANS, A., RAPOPORT, J. L. et al. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *J. Neurosci.* **28** 3586–3594.
SLATER, A. (1997). Can measures of infant habituation predict later intellectual ability? *Arch. Dis. Child.* **77** 474–476.
SMYSER, T. A., SMYSER, C. D., ROGERS, C. E., GILLESPIE, S. K., INDER, T. E. and NEIL, J. J. (2016). Cortical gray and adjacent white matter demonstrate synchronous maturation in very preterm infants. *Cereb. Cortex* **26** 3370–3378.
STILES, J. and JERNIGAN, T. (2010). The basics of brain development. *Neuropsychol. Rev.* **20** 327–348.
TAVAKOLI, S., PIGOLI, D., ASTON, J. A. and COLEMAN, J. (2016). Spatial modeling of object data: Analysing dialect sound variations across the UK. Preprint. Available at arXiv:1610.10040.
WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Ann. Rev. Stat. Appl.* **3** 257–295.
WOLFF, J. J., GU, H., GERIG, G., ELISON, J. T., STYNER, M., GOUTTARD, S., BOTTERON, K. N., DAGER, S. R., DAWSON, G., ESTES, A. M. et al. (2012). Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *Am. J. Psychiatr.* **169** 589–600.
XIAO, Z., QIU, T., KE, X., XIAO, X., XIAO, T., LIANG, F., ZOU, B., HUANG, H., FANG, H., CHU, K. et al. (2014). Autism spectrum disorder as early neurodevelopmental disorder: Evidence from the brain imaging abnormalities in 2–3 years old toddlers. *J. Autism Dev. Disord.* **44** 1633–1640.
YUAN, Y., ZHU, H., LIN, W. and MARRON, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 697–719. MR2965956
ZHOU, L., HUANG, J. Z. and CARROLL, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601–619. MR2443178

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 420–443
https://doi.org/10.1214/18-AOAS1196

THE ROLE OF MASTERY LEARNING IN AN INTELLIGENT TUTORING SYSTEM: PRINCIPAL STRATIFICATION ON A LATENT VARIABLE[1]

BY ADAM C. SALES AND JOHN F. PANE

University of Texas and RAND Corporation

Students in Algebra I classrooms typically learn at different rates and struggle at different points in the curriculum—a common challenge for math teachers. Cognitive Tutor Algebra I (CTA1), an educational computer program, addresses such student heterogeneity via what they term "mastery learning," where students progress from one section of the curriculum to the next by demonstrating appropriate "mastery" at each stage. However, when students are unable to master a section's skills even after trying many problems, they are automatically promoted to the next section anyway. Does promotion without mastery impair the program's effectiveness?

At least in certain domains, CTA1 was recently shown to improve student learning on average in a randomized effectiveness study. This paper uses student log data from that study in a continuous principal stratification model to estimate the relationship between students' potential mastery and the CTA1 treatment effect. In contrast to extant principal stratification applications, a student's propensity to master worked sections here is never directly observed. Consequently we embed an item-response model, which measures students' potential mastery, within the larger principal stratification model. We find that the tutor may, in fact, be *more* effective for students who are more frequently promoted (despite unsuccessfully completing sections of the material). However, since these students are distinctive in their educational strength (as well as in other respects), it remains unclear whether this enhanced effectiveness can be directly attributed to aspects of the mastery learning program.

REFERENCES

ANDERSON, J. R., BOYLE, C. F. and REISER, B. J. (1985). Intelligent tutoring systems. *Science* **228** 456–462.

ANDERSON, J. R., CORBETT, A. T., KOEDINGER, K. R. and PELLETIER, R. (1995). Cognitive tutors: Lessons learned. *J. Learn. Sci.* **4** 167–207.

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.

BECK, J. E. and GONG, Y. (2013). Wheel-spinning: Students who fail to master a skill. In *International Conference on Artificial Intelligence in Education* 431–440. Springer, Berlin.

BLOOM, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional Education Laboratory for the Carolinas and Virginia, topical papers and reprints, number 1. *Eval. Comment* **1** n2.

Key words and phrases. Causal inference, principal stratification, item response theory, latent variables, Bayesian, educational technology.

BOWERS, J., FREDRICKSON, M. and HANSEN, B. (2017). RItools: Randomization inference tools (development version). R package version 0.2-0. Available at https://github.com/markmfredrickson/RItools.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

DE BOECK, P. and WILSON, M., eds. (2013). *Explanatory Item Response Models*: *A Generalized Linear and Nonlinear Approach*. Springer, New York. MR2083193

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

EMBRETSON, S. E. and REISE, S. P. (2013). *Item Response Theory for Psychologists*. Psychology Press, London.

FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016a) Principal stratification in the twilight zone: Weakly separated components in finite mixture models. Preprint. Available at arXiv:1602.06595.

FELLER, A., GRINDAL, T., MIRATRIX, L. and PAGE, L. C. (2016b). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Ann. Appl. Stat.* **10** 1245–1285. MR3553224

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

GELMAN, A. and HILL, J. (2006). *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. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, *Vol.* 2, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

GILBERT, P. B. and HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262

GRIFFIN, B. A., MCCAFFREY, D. F. and MORRAL, A. R. (2008). An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. *Ann. Appl. Stat.* **2** 1034–1055. MR2516803

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

ISRANI, A., SALES, A. C. and PANE, J. F. (2018). Mastery learning in practice: A (mostly) descriptive analysis of log data from the cognitive tutor algebra I effectiveness trial. Preprint. Available at arXiv:1802.08616.

JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. MR2463484

KALTON, G. (1968). Standardization: A technique to control for extraneous variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **17** 118–136. MR0234599

KULIK, C.-L. C., KULIK, J. A. and BANGERT-DROWNS, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Rev. Educ. Res.* **60** 265–299.

LEVY, R., MISLEVY, R. J. and SINHARAY, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Appl. Psychol. Meas.* **33** 519–537.

LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. MR3456358

LITTLE, R. J. A. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data*, *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1925014

MATTEI, A., LI, F. and MEALLI, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Ann. Appl. Stat.* **7** 2336–2360. MR3161725

MIRATRIX, L., FUREY, J., FELLER, A., GRINDAL, T. and PAGE, L. C. (2017). Bounding, an accessible method for estimating principal causal effects, examined and explained. *J. Res. Educ. Eff.* **11** 133–162.

NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Rocz. Nauk Rol.* **10** 1–51.

NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. MR2847972

PAGE, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *J. Res. Educ. Eff.* **5** 215–244.

PANE, J. F., GRIFFIN, B. A., MCCAFFREY, D. F. and KARAM, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educ. Eval. Policy Anal.* **36** 127–144.

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

RASCH, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Nielson and Lydiche, Copenhagen.

RICHARDSON, T. S., EVANS, R. J. and ROBINS, J. M. (2011). Transparent parametrizations of models for potential outcomes. In *Bayesian Statistics* 9 569–610. Oxford Univ. Press, Oxford. MR3204019

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test." *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1981). Estimation in parallel randomized experiments. *J. Educ. Stat.* **6** 377–401.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SALES, A. C. and PANE, J. F. (2019). Supplement to "The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable." DOI:10.1214/18-AOAS1196SUPP.

SALES, A. C., WILKS, A. and PANE, J. F. (2016). Student usage predicts treatment effect heterogeneity in the cognitive tutor algebra I program. In *Proceedings of the 9th International Conference on Educational Data Mining. International Educational Data Mining Society* 207–214.

SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **106** 1331–1344. MR2896839

STAN DEVELOPMENT TEAM (2016). RStan: The R interface to Stan. R package version 2.14.1. Available at http://mc-stan.org/.

STEKHOVEN, D. J. and BUEHLMANN, P. (2012). Missforest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** 112–118.

STERNE, J. A. C., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. and CARPENTER, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BJM* **338** b2393.

VAN DER LINDEN, W. J. and HAMBLETON, R. K., eds. (2013). *Handbook of Modern Item Response Theory*. Springer, New York. MR1601043

YEN, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *J. Educ. Meas.* **30** 187–213.

ZHU, X. and STONE, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *J. Educ. Meas.* **48** 81–97.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 444–465
https://doi.org/10.1214/18-AOAS1198

CAPTURING HETEROGENEITY OF COVARIATE EFFECTS IN HIDDEN SUBPOPULATIONS IN THE PRESENCE OF CENSORING AND LARGE NUMBER OF COVARIATES

BY FARHAD SHOKOOHI*, ABBAS KHALILI*,[1], MASOUD ASGHARIAN*,[2] AND SHILI LIN†,[3]

*McGill University** *and Ohio State University*†

The advent of modern technology has led to a surge of high-dimensional data in biology and health sciences such as genomics, epigenomics and medicine. The high-grade serous ovarian cancer (HGS-OvCa) data reported by The Cancer Genome Atlas (TCGA) Research Network is one example. The TCGA and other research groups have analyzed several aspects of these data. Here we study the relationship between Disease Free Time (DFT) after surgery among ovarian cancer patients and their DNA methylation profiles of genomic features. Such studies pose additional challenges beyond the typical big data problem due to population substructure and censoring. Despite the availability of several methods for analyzing time-to-event data with a large number of covariates but a small sample size, there is no method available to date that accommodates the additional feature of heterogeneity. To this end, we propose a regularized framework based on the finite mixture of accelerated failure time model to capture intangible heterogeneity due to population substructure and to account for censoring simultaneously. We study the properties of the proposed framework both theoretically and numerically. Our data analysis indicates the existence of heterogeneity in the HGS-OvCa data, with one component of the mixture capturing a more aggressive form of the disease, and the second component capturing a less aggressive form. In particular, the second component portrays a significant positive relationship between methylation and DFT for BRCA1. By further unearthing the negative relationship between expression and methylation for this gene, one may provide a biologically reasonable explanation that sheds light on the relationship between DNA methylation, gene expression and mutation.

REFERENCES

BOLTON, K. L., CHENEVIX-TRENCH, G., GOH, C., SADETZKI, S., RAMUS, S. J., KARLAN, B. Y. et al. (2012). Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA* **307** 382–389.

CAI, J., FAN, J., LI, R. and ZHOU, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92** 303–316. MR2201361

CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B. E., SUMER, S. O., AKSOY, B. A. et al. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2** 401.

Key words and phrases. DNA methylation, ovarian cancer, finite mixture of AFT model, penalized regression, right censoring.

CHEN, J. and TAN, X. (2009). Inference for multivariate normal mixtures. *J. Multivariate Anal.* **100** 1367–1383. MR2514135

EARP, M. A. and CUNNINGHAM, J. M. (2015). DNA methylation changes in epithelial ovarian cancer histotypes. *Genomics* **106** 311–321.

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

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. MR1671590

GAO, J., AKSOY, B. A., DOGRUSOZ, U., DRESDNER, G., GROSS, B., SUMER, S. O. et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6** pl1.

GILBRIDE, T. J., ALLENBY, G. M. and BRAZELL, J. D. (2006). Models for heterogeneous variable selection. *J. Mark. Res.* **43** 420–430.

HAN, S. W., CHEN, G., CHEON, M.-S. and ZHONG, H. (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *J. Amer. Statist. Assoc.* **111** 1004–1019. MR3561925

HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275

KASAHARA, H. and SHIMOTSU, K. (2015). Testing the number of components in normal mixture regression models. *J. Amer. Statist. Assoc.* **110** 1632–1645. MR3449060

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662

KONG, B., LV, Z.-D., WANG, Y., JIN, L.-Y., DING, L. and YANG, Z.-C. (2015). Down-regulation of BRMS1 by DNA hypermethylation and its association with metastatic progression in triple-negative breast cancer. *Int. J. Clin. Exp. Pathol.* **8** 11076–11083.

KOUKOURA, O., SPANDIDOS, D. A., DAPONTE, A. and SIFAKIS, S. (2014). DNA methylation profiles in ovarian cancer: Implication in diagnosis and therapy (review). *Mol. Med. Rep.* **10** 3–9.

KWON, M. J. and SHIN, Y. K. (2011). Epigenetic regulation of cancer-associated genes in ovarian cancer. *Int. J. Mol. Sci.* **12** 983–1008.

KWONG, J., LEE, J.-Y., WONG, K.-K., ZHOU, X., WONG, D. T. W., LO, K.-W. et al. (2006). Candidate tumor-suppressor gene DLEC1 is frequently downregulated by promoter hypermethylation and histone hypoacetylation in human epithelial ovarian cancer. *Neoplasia* **8** 268–278.

LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1940115

LEE, K. H., CHAKRABORTY, S. and SUN, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *Int. J. Biostat.* **7** Art. 21. MR2787411

LIU, J., ZHANG, R., ZHAO, W. and LV, Y. (2015). Variable selection in semiparametric hazard regression for multivariate survival data. *J. Multivariate Anal.* **142** 26–40. MR3412736

LU, Z.-H. (2009). Covariate selection in mixture models with the censored response variable. *Comput. Statist. Data Anal.* **53** 2710–2723. MR2665920

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

MCLACHLAN, G. J. and MCGIFFIN, D. C. (1994). On the role of finite mixture models in survival analysis. *Stat. Methods Med. Res.* **3** 211–226.

MCLACHLAN, G. and PEEL, D. (2004). *Finite Mixture Models*. Wiley, New York.

SCHÖNDORF, T., EBERT, M. P., HOFFMANN, J., BECKER, M., MOSER, N., PUR, Ş. et al. (2016). Hypermethylation of the PTEN gene in ovarian cancer cell lines. *Cancer Lett.* **207** 215–220.

SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268.

SHOKOOHI, F., KHALILI, A., ASGHARIAN, M. and LIN, S. (2019). Supplement to "Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates." DOI:10.1214/18-AOAS1198SUPP.

SOHN, I., KIM, J., JUNG, S.-H. and PARK, C. (2009). Gradient Lasso for Cox proportional hazards model. *Bioinformatics* **25** 1775–1781.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). L_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

THE CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609–615.

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

TIBSHIRANI, R. (1997). The Lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

VOLINSKY, C. T. and RAFTERY, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56** 256–262.

YANG, D., KHAN, S., SUN, Y., HESS, K., SHMULEVICH, I., SOOD, A. K. et al. (2011). Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* **306** 1557–1565.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, F., DING, J. and LIN, S. (2017). Testing for associations of opposite directionality in a heterogeneous population. *Statist. Biosci.* **9** 137–159.

ZHU, H.-T. and ZHANG, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 3–16. MR2035755

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 466–491
https://doi.org/10.1214/18-AOAS1202

DEVELOPMENT OF A COMMON PATIENT ASSESSMENT SCALE ACROSS THE CONTINUUM OF CARE: A NESTED MULTIPLE IMPUTATION APPROACH[1]

BY CHENYANG GU AND ROEE GUTMAN

Harvard University and Brown University

Evaluating and tracking patients' functional status through the post-acute care continuum requires a common instrument. However, different post-acute service providers such as nursing homes, inpatient rehabilitation facilities and home health agencies rely on different instruments to evaluate patients' functional status. These instruments assess similar functional status domains, but they comprise different activities, rating scales and scoring instructions. These differences hinder the comparison of patients' assessments across health care settings. We propose a two-step procedure that combines nested multiple imputation with the multivariate ordinal probit (MVOP) model to obtain a common patient assessment scale across the post-acute care continuum. Our procedure imputes the unmeasured assessments at multiple assessment dates and enables evaluation and comparison of the rates of functional improvement experienced by patients treated in different health care settings using a common measure. To generate multiple imputations of the unmeasured assessments using the MVOP model, a likelihood-based approach that combines the EM algorithm and the bootstrap method as well as a fully Bayesian approach using the data augmentation algorithm are developed. Using a dataset on patients who suffered a stroke, we simulate missing assessments and compare the MVOP model to existing methods for imputing incomplete multivariate ordinal variables. We show that, for all of the estimands considered, and in most of the experimental conditions that were examined, the MVOP model appears to be superior. The proposed procedure is then applied to patients who suffered a stroke and were released from rehabilitation facilities either to skilled nursing facilities or to their homes.

REFERENCES

ABAYOMI, K., GELMAN, A. and LEVY, M. (2008). Diagnostics for multivariate imputations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **57** 273–291. MR2440009

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ANDRIDGE, R. R. LITTLE, R. J. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **78** 40–64.

ASHFORD, J. R. and SOWDEN, R. R. (1970). Multi-variate probit analysis. *Biometrics* **26** 535–546.

BURGETTE, L. F. and REITER, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *Am. J. Epidemiol.* **172** 1070–1076.

Key words and phrases. Data augmentation, EM algorithm, missing data, nested multiple imputation, multivariate ordinal probit model, slice sampler.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P., RIDDELL, A. et al. (2016). Stan: A probabilistic programming language. *J. Stat. Softw.* **20** 1–37.

CHEN, M.-H. and DEY, D. K. (2000). A unified Bayesian approach for analyzing correlated ordinal response data. *Braz. J. Probab. Stat.* **14** 87–111. MR1838453

CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006). *Statistical Matching*: *Theory and Practice*. Wiley, Chichester. MR2268833

DAMIEN, P. and WALKER, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *J. Comput. Graph. Statist.* **10** 206–215. MR1939697

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DORANS, N. J., POMMERICH, M. and HOLLAND, P. W. (2007). *Linking and Aligning Scores and Scales*. Springer, New York.

ENDERS, C. K., KELLER, B. T. and LEVY, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychol. Methods* **23** 298–317.

FISCHER, H. F., TRITT, K., KLAPP, B. F. and FLIEGE, H. (2011). How to compare scores from different depression scales: Equating the Patient Health Questionnaire (PHQ) and the ICD-10-Symptom Rating (ISR) using Item Response Theory. *Int. J. Methods Psychiatr. Res.* **20** 203–214.

GAGE, B., CONSTANTINE, R., AGGARWAL, J., MORLEY, M., KURLANTZICK, V., BERNARD, S. et al. (2012). The development and testing of the Continuity Assessment Record and Evaluation. (CARE) item set: Final report on the development of the CARE item set.

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

GELMAN, A., VAN MECHELEN, I., VERBEKE, G., HEITJAN, D. F. and MEULDERS, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61** 74–85. MR2135847

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* **1** 141–149.

GEWEKE, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing Science and Statistics*: *Proceedings of the* 23*rd Symposium on the Interface* 571–578. Interface Foundation of North America, Fairfax, VA.

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

GOODMAN, L. A. and KRUSKAL, W. H. (1979). *Measures of Association for Cross Classifications*. *Springer Series in Statistics* **1**. Springer, New York. MR0553108

GU, C. and GUTMAN, R. (2017). Combining item response theory with multiple imputation to equate health assessment questionnaires. *Biometrics* **73** 990–998. MR3713132

GU, C. and GUTMAN, R. (2019). Supplement to "Development of a common patient assessment scale across the continuum of care: A nested multiple imputation approach." DOI:10.1214/18-AOAS1202SUPP.

GUERRERO, V. M. and JOHNSON, R. A. (1982). Use of the Box–Cox transformation with binary response models. *Biometrika* **69** 309–314. MR0671968

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2015). Graphical models for ordinal data. *J. Comput. Graph. Statist.* **24** 183–204. MR3328253

HAREL, O. (2003). Strategies for data analysis with two types of missing values. PhD thesis, Pennsylvania State Univ., State College, PA.

HE, Y. and ZASLAVSKY, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Stat. Med.* **31** 1–18. MR2868986

HEITJAN, D. F., LANDIS and RICHARD, J. (1994). Assessing secular trends in blood pressure: A multiple-imputation approach. *J. Amer. Statist. Assoc.* **89** 750–759.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive *p*-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154

HOLMES, C. C. and HELD, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1** 145–168. MR2227368

JELIAZKOV, I., GRAVES, J. and KUTZBACH, M. (2008). Fitting and comparison of models for multivariate ordinal outcomes. *Adv. Econom.* **23** 115–156.

KOLEN, M. J. and BRENNAN, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd ed. Springer, New York. MR3156933

LAWRENCE, E., LIU, C., BINGHAM, D. and NAIR, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics* **50** 182–191. MR2439877

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081

LI, Y. and SCHAFER, D. W. (2008). Likelihood analysis of the multivariate ordinal probit regression model for repeated ordinal responses. *Comput. Statist. Data Anal.* **52** 3474–3492. MR2427359

LI, C.-Y., ROMERO, S., SIMPSON, K. N., BONILHA, H. S., SIMPSON, A. N., HONG, I. and VELOZO, C. A. (2017). Linking existing instruments to develop a continuum of care measure: Accuracy comparison using function-related group classification. *Qual. Life Res.* 1–10.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85** 755–770. MR1666758

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR3223057

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

OLSSON, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44** 443–460. MR0554892

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

R CORE TEAM (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1994). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

RUBIN, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. MR2055518

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability* **72**. Chapman & Hall, London. MR1692799

SHEN, Z. (2000). Nested multiple imputations. PhD thesis, Harvard Univ., Cambridge, MA.

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38** 499–521.

SI, Y., REITER, J. P. and HILLYGUS, D. S. (2016). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. *Ann. Appl. Stat.* **10** 118–143. MR3480490

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

TEN KLOOSTER, P. M., VOSHAAR, M. A. H. O., GANDEK, B., ROSE, M., BJORNER, J. B., TAAL, E., GLAS, C. A. W., VAN RIEL, P. L. C. M. and VAN DE LAAR, M. A. F. J. (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. *Health Qual Life Outcomes* **11** 199.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16** 219–242. MR2371007

VARIN, C. and CZADO, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics* **11** 127–138.

VELOZO, C. A., BYERS, K. L. and JOSEPH, B. R. (2007). Translating measures across the continuum of care: Using Rasch analysis to create a crosswalk between the functional independence measure and the minimum data set. *J. Rehabil. Res. Dev.* **44** 467.

VERMUNT, J. K., VAN GINKEL, J. R., DER ARK, V., ANDRIES, L. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Method.* **38** 369–397.

VON DAVIER, A. A., ed. (2011). *Statistical Models for Test Equating, Scaling, and Linking. Statistics for Social and Behavioral Sciences*. Springer, New York. MR2798186

WEI, C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WYSOCKI, A., THOMAS, K. S. and MOR, V. (2015). Functional improvement among short-stay nursing home residents in the MDS 3.0. *J. Am. Med. Dir. Assoc.* **16** 470–474.

YUCEL, R. M., HE, Y. and ZASLAVSKY, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Stat. Med.* **30** 3447–3460. MR2861625

ZHANG, X., BOSCARDIN, W. J. and BELIN, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *J. Comput. Graph. Statist.* **15** 880–896. MR2297633

ZHANG, X., LI, Q., CROPSEY, K., YANG, X., ZHANG, K. and BELIN, T. (2017). A multiple imputation method for incomplete correlated ordinal data using multivariate probit models. *Comm. Statist. Simulation Comput.* **46** 2360–2375. MR3625287

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 492–519
https://doi.org/10.1214/18-AOAS1203

A BAYESIAN MALLOWS APPROACH TO NONTRANSITIVE PAIR COMPARISON DATA: HOW HUMAN ARE SOUNDS?

BY MARTA CRISPINO[*,1], ELJA ARJAS[†,‡], VALERIA VITELLI[‡], NATASHA BARRETT[§] AND ARNOLDO FRIGESSI[‡,¶]

Inria Grenoble[*], University of Helsinki[†], University of Oslo[‡], Norwegian State Academy for Music in Oslo[§], Oslo University Hospital[¶]

We are interested in learning how listeners perceive sounds as having human origins. An experiment was performed with a series of electronically synthesized sounds, and listeners were asked to compare them in pairs. We propose a Bayesian probabilistic method to learn individual preferences from nontransitive pairwise comparison data, as happens when one (or more) individual preferences in the data contradicts what is implied by the others. We build a Bayesian Mallows model in order to handle nontransitive data, with a latent layer of uncertainty which captures the generation of preference misreporting. We then develop a mixture extension of the Mallows model, able to learn individual preferences in a heterogeneous population. The results of our analysis of the musicology experiment are of interest to electroacoustic composers and sound designers, and to the audio industry in general, whose aim is to understand how computer generated sounds can be produced in order to sound more human.

REFERENCES

AGRESTI, A. (1996). *Categorical Data Analysis*, Wiley, New York.

BARRETT, N. and CRISPINO, M. (2018). The impact of 3-D sound spatialisation on listeners' understanding of human agency in acousmatic music. *J. New Music Res.* 1–17.

BIERNACKI, C. and JACQUES, J. (2013). A generative model for rank data based on insertion sort algorithm. *Comput. Statist. Data Anal.* **58** 162–176. MR2997933

BÖCKENHOLT, U. (1988). A logistic representation of multivariate paired-comparison models. *J. Math. Psych.* **32** 44–63. MR0935673

BÖCKENHOLT, U. (2001). Hierarchical modeling of paired comparison data. *Psychol. Methods* **6** 49–66.

BÖCKENHOLT, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika* **71** 615–629. MR2312235

BÖCKENHOLT, U. and TSAI, R.-C. (2001). Individual differences in paired comparison data. *Br. J. Math. Stat. Psychol.* **54** 265–277.

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925

CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181. MR3262549

Key words and phrases. Nontransitive pairwise comparisons, ranking, Mallows model, Bayesian preference learning, recommender systems, musicology, acousmatic experiment.

CAYLEY, A. (1849). LXXVII. Note on the theory of permutations. *Philos. Mag. Ser.* 3 **34** 527–529.
CRISPINO, M. and FRIGESSI, A. (2018). The hierarchical Bradley–Terry model. *Draft*.
CRISPINO, M., ARJAS, E., VITELLI, V., BARRETT, N. and FRIGESSI, A. (2019). Supplement to "A Bayesian Mallows approach to nontransitive pair comparison data: How human are sounds?" DOI:10.1214/18-AOAS1203SUPP.
DAVIDSON, R. R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *J. Amer. Statist. Assoc.* **65** 317–328.
DE BORDA, J. C. (1781). *Mémoire sur les Élections Au Scrutin*, *Histoire de L'Académie Royale des Sciences*. Paris, France.
DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069
DING, W., ISHWAR, P. and SALIGRAMA, V. (2015). Learning mixed membership Mallows models from pairwise comparisons. Preprint. Available at ArXiv:1504.00757.
DITTRICH, R., HATZINGER, R. and KATZENBEISSER, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *J. Roy. Statist. Soc. Ser. C* **47** 511–525.
DITTRICH, R., HATZINGER, R. and KATZENBEISSER, W. (2002). Modelling dependencies in paired comparison data: A log-linear approach. *Comput. Statist. Data Anal.* **40** 39–57. MR1921121
DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank aggregation methods for the Web. In *Proceedings of the* 10*th International Conference on World Wide Web* 613–622. ACM.
FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48** 359–369. MR0876847
FORD, L. R. JR. (1957). Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* **64** 28–33. MR0097876
FRANCIS, B., DITTRICH, R. and HATZINGER, R. (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do Europeans get their scientific knowledge? *Ann. Appl. Stat.* **4** 2181–2202. MR2829952
GORMLEY, I. C. and MURPHY, T. B. (2006). Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A* **169** 361–379. MR2225548
GROND, F. and BERGER, J. (2011). Parameter Mapping Sonification. In *The Sonification Handbook* (T. Hermann, A. D. Hunt and J. Neuhoff, eds.) 363–398. Logos Publishing House, Berlin.
IRUROZKI, E., CALVO, B. and LOZANO, A. (2014). Sampling and learning the Mallows and generalized Mallows models under the Hamming distance. Technical report, Univ. del Pais Vasco, San Sebastian, Spain.
IRUROZKI, E., CALVO, B. and LOZANO, J. A. (2018). Sampling and learning Mallows and generalized Mallows models under the Cayley distance. *Methodol. Comput. Appl. Probab.* **20** 1–35. MR3760337
JACQUES, J. and BIERNACKI, C. (2014). Model-based clustering for multivariate partial ranking data. *J. Statist. Plann. Inference* **149** 201–217. MR3199905
KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.
KENYON-MATHIEU, C. and SCHUDY, W. (2007). How to rank with few errors. In *STOC'*07—*Proceedings of the* 39*th Annual ACM Symposium on Theory of Computing* 95–103. ACM, New York. MR2402432
LIU, Q., CRISPINO, M., SCHEEL, I., VITELLI, V. and FRIGESSI, A. (2018). Model-based learning from preference data. *Ann. Rev. Stat. Appl.*. To appear.
LU, T. and BOUTILIER, C. (2014). Effective sampling and learning for Mallows models with pairwise-preference data. *J. Mach. Learn. Res.* **15** 3963-4009. MR3317212
LUCE, R. D. (1959). *Individual Choice Behavior*: *A Theoretical Analysis*. Wiley, New York. MR0108411

MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267
MARQUIS OF CONDORCET, M. J. A. N. D. C. (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. *Paris*: *De L'imprimerie Royale*.
MEILĂ, M. and CHEN, H. (2010). Dirichlet process mixtures of generalized Mallows models. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*-10) 358–367. AUAI Press, Corvallis, OR, USA.
MUKHERJEE, S. (2016). Estimation in exponential families on permutations. *Ann. Statist.* **44** 853–875. MR3476619
MURPHY, T. B. and MARTIN, D. (2003). Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.* **41** 645–655. MR1973732
NEGAHBAN, S., OH, S. and SHAH, D. (2012). Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems* 2474–2482.
OLLEN, J. E. (2006). A criterion-related validity test of selected indicators of musical sophistication using expert ratings Ph.D. thesis, Ohio State Univ., Columbus, OH.
PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. MR0391338
RAJKUMAR, A., GHOSHAL, S., LIM, L.-H. and AGARWAL, S. (2015). Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *ICML* 665–673.
RAO, P. V. and KUPPER, L. L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *J. Amer. Statist. Assoc.* **62** 194–204. MR0217963
SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* **15** 72–101.
SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380
STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293
THURSTONE, L. L. (1927). A law of comparative judgment. *Psychol. Rev.* **34** 273.
TVERSKY, A. (1969). Intransitivity of preferences. *Preference*, *Belief*, *and Similarity* 433-461.
VITELLI, V., SØRENSEN, Ø., CRISPINO, M., FRIGESSI, A. and ARJAS, E. (2018). Probabilistic preference learning with the Mallows rank model. *J. Mach. Learn. Res.* **18(158)** 1–49. MR3813807
VOLKOVS, M. N. and ZEMEL, R. S. (2014). New learning methods for supervised and unsupervised preference aggregation. *J. Mach. Learn. Res.* **15** 1135–1176. MR3195341
YAN, T. (2016). Ranking in the generalized Bradley–Terry models when the strong connection condition fails. *Comm. Statist. Theory Methods* **45** 340–353. MR3447919
ZERMELO, E. (1929). Die Berechnung der Turnier–Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.* **29** 436–460. MR1545015

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 520–547
https://doi.org/10.1214/18-AOAS1206

CAUSAL INFERENCE IN THE CONTEXT OF AN ERROR PRONE EXPOSURE: AIR POLLUTION AND MORTALITY[1]

BY XIAO WU[*], DANIELLE BRAUN[*], MARIANTHI-ANNA KIOUMOURTZOGLOU[†], CHRISTINE CHOIRAT[*], QIAN DI[*] AND FRANCESCA DOMINICI[*]

Harvard T.H. Chan School of Public Health[*] and Columbia University Mailman School of Public Health[†]

We propose a new approach for estimating causal effects when the exposure is measured with error and confounding adjustment is performed via a generalized propensity score (GPS). Using validation data, we propose a regression calibration (RC)-based adjustment for a continuous error-prone exposure combined with GPS to adjust for confounding (RC-GPS). The outcome analysis is conducted after transforming the corrected continuous exposure into a categorical exposure. We consider confounding adjustment in the context of GPS subclassification, inverse probability treatment weighting (IPTW) and matching. In simulations with varying degrees of exposure error and confounding bias, RC-GPS eliminates bias from exposure error and confounding compared to standard approaches that rely on the error-prone exposure. We applied RC-GPS to a rich data platform to estimate the causal effect of long-term exposure to fine particles ($PM_{2.5}$) on mortality in New England for the period from 2000 to 2012. The main study consists of 2202 zip codes covered by 217,660 1 km $\times$ 1 km grid cells with yearly mortality rates, yearly $PM_{2.5}$ averages estimated from a spatio-temporal model (error-prone exposure) and several potential confounders. The internal validation study includes a subset of 83 1 km $\times$ 1 km grid cells within 75 zip codes from the main study with error-free yearly $PM_{2.5}$ exposures obtained from monitor stations. Under assumptions of noninterference and weak unconfoundedness, using matching we found that exposure to moderate levels of $PM_{2.5}$ ($8 < PM_{2.5} \leq 10\ \mu g/m^3$) causes a 2.8% (95% CI: 0.6%, 3.6%) increase in all-cause mortality compared to low exposure ($PM_{2.5} \leq 8\ \mu g/m^3$).

REFERENCES

ABADIE, A. and IMBENS, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557. MR2468559

ABADIE, A. and IMBENS, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84** 781–807. MR3481379

ALEXEEFF, S. E., CARROLL, R. J. and COULL, B. (2016). Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics* **17** 377–389. MR3516007

Key words and phrases. Measurement error, generalized propensity scores, observational study, air pollution, environmental epidemiology, causal inference.

BABANEZHAD, M., VANSTEELANDT, S. and GOETGHEBEUR, E. (2010). Comparison of causal effect estimators under exposure misclassification. *J. Statist. Plann. Inference* **140** 1306–1319. MR2581132

BACCINI, M., MATTEI, A., MEALLI, F., BERTAZZI, P. A. and CARUGNO, M. (2017). Assessing the short term impact of air pollution on mortality: A matching approach. *Environ. Health* **16** 7.

BEELEN, R., RAASCHOU-NIELSEN, O., STAFOGGIA, M., ANDERSEN, Z. J., WEINMAYR, G., HOFFMANN, B., WOLF, K., SAMOLI, E., FISCHER, P. and NIEUWENHUIJSEN, M. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet* **383** 785–795.

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (2012). Resampling fewer than n observations: Gains, losses, and remedies for losses. In *Selected Works of Willem van Zwet* 267–297. Springer, Berlin.

BRAUN, D., GORFINE, M., PARMIGIANI, G., ARVOLD, N. D., DOMINICI, F. and ZIGLER, C. (2017). Propensity scores with misclassified treatment assignment: A likelihood-based adjustment. *Biostatistics* **18** 695–710. MR3799597

BURTON, R. M., SUH, H. H. and KOUTRAKIS, P. (1996). Spatial variation in particulate concentrations within metropolitan Philadelphia. *Environ. Sci. Technol.* **30** 400–407.

CARROLL, R. J. and STEFANSKI, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85** 652–663. MR1138349

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; CRC Press, London. MR0095561

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DEHEJIA, R. H. and WAHBA, S. (1998). Propensity score matching methods for non-experimental casual studies.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

DI, Q., KOUTRAKIS, P. and SCHWARTZ, J. (2016). A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression. *Atmos. Environ.* **131** 390–399.

DI, Q., KLOOG, I., KOUTRAKIS, P., LYAPUSTIN, A., WANG, Y. and SCHWARTZ, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **50** 4712–4721.

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., ZEGER, S. L. and SAMET, J. M. (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* **1** 157–175.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903

FULLER, W. A. (2009). *Measurement Error Models*. Wiley Interscience, Hoboken, NJ. Reprint of the 1987 original, *Wiley-Interscience Paperback Series*. MR2301581

GAIL, M. H., WU, J., WANG, M. et al. (2016). Calibration and seasonal adjustment for matched case-control studies of vitamin D and cancer. *Stat. Med.* **35** 2133–2148. MR3513504

GRYPARIS, A., PACIOREK, C. J., ZEKA, A., SCHWARTZ, J. and COULL, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10** 258–274.

HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol. Methods* **15** 234–249.

HART, J. E., LIAO, X., HONG, B., PUETT, R. C., YANOSKY, J. D., SUH, H., KIOUMOURTZOGLOU, M.-A., SPIEGELMAN, D. and LADEN, F. (2015). The association of long-term exposure to $PM_{2.5}$ on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ. Health* **14** 38.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMBENS, G. W. (2000). The role of the propensity score in estimating dose–response functions. *Biometrika* **87** 706–710. MR1789821

KIOUMOURTZOGLOU, M.-A., SPIEGELMAN, D., SZPIRO, A. A., SHEPPARD, L., KAUFMAN, J. D., YANOSKY, J. D., WILLIAMS, R., LADEN, F., HONG, B. and SUH, H. (2014). Exposure measurement error in $PM_{2.5}$ health effects studies: A pooled analysis of eight personal exposure validation studies. *Environ. Health* **13** 2.

KIOUMOURTZOGLOU, M.-A., SCHWARTZ, J., JAMES, P., DOMINICI, F. and ZANOBETTI, A. (2016). PM2.5 and mortality in 207 US cities: Modification by temperature and city characteristics. *Epidemiology* **27** 221–227.

LECHNER, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* 43–58. Springer, Berlin.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

PIERCE, D. A. and KELLERER, A. M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika* **91** 863–876. MR2126038

RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. and SCHNEEWEISS, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24** 401–409.

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, Philaldelphia, PA.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

SARNAT, S. E., KLEIN, M., SARNAT, J. A., FLANDERS, W. D., WALLER, L. A., MULHOLLAND, J. A., RUSSELL, A. G. and TOLBERT, P. E. (2010). An examination of exposure measurement error from air pollutant spatial variability in time-series studies. *J. Expo. Sci. Environ. Epidemiol.* **20** 135–146.

SHI, L., ZANOBETTI, A., KLOOG, I., COULL, B. A., KOUTRAKIS, P., MELLY, S. J. and SCHWARTZ, J. D. (2016). Low-concentration PM2.5 and mortality: Estimating acute and chronic effects in a population-based study. *Environ. Health Perspect.* **124** 46.

SZPIRO, A. A., SHEPPARD, L. and LUMLEY, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12** 610–623.

USEPA (2012). US Environmental Protection Agency. National Ambient Air Quality Standards (NAAQS) table. Available at https://www.epa.gov/criteria-air-pollutants/naaqs-table.

VAN ROOSBROECK, S., LI, R., HOEK, G., LEBRET, E., BRUNEKREEF, B. and SPIEGELMAN, D. (2008). Traffic-related outdoor air pollution and respiratory symptoms in children: The impact of adjustment for exposure measurement error. *Epidemiology* **19** 409–416.

VAUGHN, B. K. (2008). Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J. *J. Educ. Meas.* **45** 94–97.

VILLENEUVE, P. J., WEICHENTHAL, S. A., CROUSE, D., MILLER, A. B., TO, T., MARTIN, R. V., VAN DONKELAAR, A., WALL, C. and BURNETT, R. T. (2015). Long-term exposure to fine particulate matter air pollution and mortality among Canadian women. *Epidemiology* **26** 536–545.

WHO (2018). World Health Organization. Air quality guidelines. Available at http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

WILSON, W. E. and SUH, H. H. (1997). Fine particles and coarse particles: Concentration relationships relevant to epidemiologic studies. *J. Air Waste Manage. Assoc.* **47** 1238–1249.

WU, X., BRAUN, D., KIOUMOURTZOGLOU, M.-A., CHOIRAT, C., DI, Q. and DOMINICI, F. (2019). Supplement to "Causal inference in the context of an error prone exposure: Air pollution and mortality." DOI:10.1214/18-AOAS1206SUPP.

YANG, S., IMBENS, G. W., CUI, Z., FARIES, D. E. and KADZIOLA, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72** 1055–1065. MR3591590

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 548–572
https://doi.org/10.1214/18-AOAS1207

MODELING BIOMARKER RATIOS WITH GAMMA DISTRIBUTED COMPONENTS[1]

BY MORITZ BERGER[*], MICHAEL WAGNER[*,†] AND MATTHIAS SCHMID[*,†]

University of Bonn[*] and German Center for Neurodegenerative Diseases[†]

We propose a regression model termed "extended GB2 model", which is designed to analyze ratios of biomarkers in epidemiological and medical research. Typical examples of biomarker ratios are given by the LDL/HDL cholesterol ratio in cardiovascular research and the amyloid-β 42/40 ratio in dementia research. Unlike regression modeling with a log-transformed response, which is often used to describe ratio outcomes in observational studies, the extended GB2 model directly links the expectation of the untransformed biomarker ratio to a set of covariates. This strategy allows for a simple interpretation of the predictor-response relationships in terms of multiplicative increases/decreases of the expected outcome, similar to Poisson and Cox regression. In the theoretical part of the paper, we derive the log-likelihood of the proposed model, analyze its properties, and provide details on confidence intervals and hypothesis testing. We will also present the results of a simulation study demonstrating the robustness of the proposed modeling approach against model misspecification. The usefulness of the method is demonstrated by two applications on the aforementioned LDL/HDL cholesterol and amyloid-β 42/40 ratios. For this, we analyze data from a cohort study on kidney disease and from a large observational database on neurodegenerative diseases.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. CRC Press, London. MR0865647

BALAKRISHNAN, N. and LAI, C.-D. (2009). *Continuous Bivariate Distributions*, 2nd ed. Springer, Dordrecht. MR2840643

BERGER, M. (2018). Supplemental files to "Modeling Biomarker Ratios with Gamma Distributed Components." Available at https://imbie.meb.uni-bonn.de/~berger/RCode_Simulations.zip.

BERGER, M. and SCHMID, M. (2019). eGB2: Fitting (Extended) GB2 Models R package version 1.0.1. DOI:10.1214/18-AOAS1207SUPPB.

BERGER, M., WAGNER, M. and SCHMID, M. (2019). Supplement to "Modeling Biomarker Ratios with Gamma Distributed Components." DOI:10.1214/18-AOAS1207SUPPA.

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach*, 2nd ed. Springer, New York. MR1919620

CABY, F., GUIHOT, A., LAMBERT-NICLOT, S., GUIGUET, M., BOUTOLLEAU, D., AGHER, R., VALANTIN, M.-A., TUBIANA, R., CALVEZ, V., MARCELIN, A.-G., CARCELAIN, G., AUTRAN, B., COSTAGLIOLA, D. and KATLAMA, C. (2016). Determinants of a low CD4/CD8 ratio

Key words and phrases. Biomarker ratio, gamma distribution, generalized beta distribution of the second kind, generalized linear model, regression.

in HIV-1-infected individuals despite long-term viral suppression. *Clin. Infect. Dis.* **62** 1297–1303.

CANDAN, Ç. and ORGUNER, U. (2013). The moment function for the ratio of correlated generalized gamma variables. *Statist. Probab. Lett.* **83** 2353–2356. MR3093825

DIGGLE, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics* **45** 1255–1258.

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–244.

ECKARDT, K. U., BÄRTHLEIN, B., BAID-AGRAWAL, S., BECK, A., BUSCH, M., EITNER, F., EKICI, A. B., FLOEGE, J., GEFELLER, O., HALLER, H., HILGE, R., HILGERS, K. F., KIELSTEIN, J. T., KRANE, V., KÖTTGEN, A., KRONENBERG, F., OEFNER, P., PROKOSCH, H. U., REIS, A., SCHMID, M., SCHAEFFNER, E., SCHULTHEISS, U. T., SEUCHTER, S. A., SITTER, T., SOMMERER, C., WALZ, G., WANNER, C., WOLF, G., ZEIER, M. and TITZE, S. (2012). The German Chronic Kidney Disease (GCKD) study: Design and methods. *Nephrol. Dial. Transplant.* **27** 1454–1460.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485

FIRTH, D. (1988). Multiplicative errors: Log-normal or gamma? *J. Roy. Statist. Soc. Ser. B* **50** 266–268. MR0964179

GANSEVOORT, R. T., CORREA-ROTTER, R., HEMMELGARN, B. R., JAFAR, T. H., HEERSPINK, H. J., MANN, J. F., MATSUSHITA, K. and WEN, C. P. (2013). Chronic kidney disease and cardiovascular risk: Epidemiology, mechanisms, and prevention. *Lancet* **27** 339–352.

HOFNER, B., MAYR, A., ROBINZONOV, N. and SCHMID, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Comput. Statist.* **29** 3–35. MR3260108

HOJSGAARD, S., HALEKOH, U. and YAN, J. (2016). geepack: Generalized Estimating Equation Package R package version 1.2-1. Available at https://cran.r-project.org/web/packages/geepack.

JACK, C. R., WISTE, H. J., WEIGAND, S. D., KNOPMAN, D. S., VEMURI, P., MIELKE, M. M., LOWE, V., SENJEM, M. L., GUNTER, J. L., MACHULDA, M. M., GREGG, B. E., PANKRATZ, V. S., ROCCA, W. A. and PETERSEN, R. C. (2015). Age, sex, and APOE $\varepsilon 4$ effects on memory, brain structure, and β-amyloid across the adult life span. *JAMA Neurol.* **72** 511–519.

KIBBLE, W. F. (1941). A two-variate gamma type distribution. *Sankhyā* **5** 137–150. MR0007218

KLEIBER, C. and KOTZ, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Interscience, Hoboken, NJ. MR1994050

KORNHUBER, J., SCHMIDTKE, K., FRÖLICH, L., PERNECZKY, R., WOLF, S., HAMPEL, H., JESSEN, F., HEUSER, I., PETERS, O., WEIH, M., JAHN, H., LUCKHAUS, C., HÜLL, M., GERTZ, H. J., SCHRÖDER, J., PANTEL, J., RIENHOFF, O., SEUCHTER, S. A., RÜTHER, E., HENN, F., MAIER, W. and WILTFANG, J. (2009). Early and differential diagnosis of dementia and mild cognitive impairment: Design and cohort baseline characteristics of the German Dementia Competence Network. *Dement. Geriatr. Cogn. Disord.* **27** 404–417.

KOYAMA, A., OKEREKE, O. I., YANG, T., BLACKER, D., SELKOE, D. J. and GRODSTEIN, F. (2012). Plasma amyloid-beta as a predictor of dementia and cognitive decline—a systematic review and meta-analysis. *Archives of Neurology* **69** 824–831.

LEE, R. Y., HOLLAND, B. S. and FLUECK, J. A. (1979). Distribution of a ratio of correlated gamma random variables. *SIAM J. Appl. Math.* **36** 304–320. MR0524504

LEWCZUK, P., LELENTAL, N., SPITZER, P., MALER, J. M. and KORNHUBER, J. (2015). Amyloid-β 42/40 cerebrospinal fluid concentration ratio in the diagnostics of Alzheimer's disease: Validation of two novel assays. *J. Alzheimer's Dis.* **43** 183–191.

LINN, S., CARROLL, M., JOHNSON, C., FULWOOD, R., KALSBEEK, W. and BRIEFEL, R. (1993). High-density lipoprotein cholesterol and alcohol consumption in US white and black adults: Data from NHANES II. *Am. J. Publ. Health* **83** 811–816.

LONG, Q., ZHANG, X., ZHAO, Y., JOHNSON, B. A. and BOSTICK, R. M. (2016). Modeling clinical outcome using multiple correlated functional biomarkers: A Bayesian approach. *Stat. Methods Med. Res.* **25** 520–537. MR3489650

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. CRC Press, London. MR3223057

MENDIS, S., PUSKA, P. and NORRVING, B., eds. (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization, Geneva.

MILLAN, J., PINTO, X., MUNOZ, A., ZUNIGA, M., RUBIES-PRAT, J., PALLARDO, L. F., MASANA, L., MANGAS, A., HERNANDEZ-MIJARES, A., GONZALEZ-SANTOS, P., ASCASO, J. F. and PEDRO-BOTET, J. (2009). Lipoprotein ratios: Physiological significance and clinical usefulness in cardiovascular prevention. *Vasc. Health Risk Manag.* **5** 757–765.

MÜLLER, H., LINDMAN, A. S., BRANTSAETER, A. L. and PEDERSEN, J. I. (2003). The serum LDL/HDL cholesterol ratio is influenced more favorably by exchanging saturated with unsaturated fat than by reducing saturated fat in the diet of women. *J. Nutr.* **133** 78–83.

NADARAJAH, S. and KOTZ, S. (2007). Jensen's bivariate gamma distribution: Ratios of components. *J. Stat. Comput. Simul.* **77** 349–358. MR2345738

NATARAJAN, S., GLICK, H., CRIQUI, M., HOROWITZ, D., LIPSITZ, S. R. and KINOSIAN, B. (2003). Cholesterol measures to identify and treat individuals at risk for coronary heart disease. *Am. J. Prev. Med.* **25** 50–57.

POTTHOFF, R. F., TUDOR, G. E., PIEPER, K. S. and HASSELBLAD, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Stat. Methods Med. Res.* **15** 213–234. MR2227446

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. Roy. Statist. Soc. Ser. C* **54** 507–554. MR2137253

SACKS, F. M., LICHTENSTEIN, A., VAN HORN, L., HARRIS, W., KRIS-ETHERTON, P. and WINSTON, M. (2006). Soy protein, isoflavones, and cardiovascular health: An American Heart Association Science Advisory for professionals from the Nutrition Committee. *Circulation* **113** 1034–1044.

SHAMAI, L., LURIX, E., SHEN, M., NOVARO, G. M., SZOMSTEIN, S., ROSENTHAL, R., HERNANDEZ, A. V. and ASHER, C. R. (2011). Association of body mass index and lipid profiles: Evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obes. Surg.* **21** 42–47.

SPERLING, R. A., KARLAWISH, J. and JOHNSON, K. A. (2013). Preclinical Alzheimer disease—the challenges ahead. *Nat. Rev. Neurol.* **9** 54–58.

SUNDRAM, K., KARUPAIAH, T. and HAYES, K. C. (2007). Stearic acid-rich interesterified fat and trans-rich fat raise the LDL/HDL ratio and plasma glucose relative to palm olein in humans. *Nutr. Metab.* **4** 3.

TITZE, S., SCHMID, M., KÖTTGEN, A., BUSCH, M., FLOEGE, J., WANNER, C., KRONENBERG, F. and ECKARDT, K. U. (2015). Disease burden and risk profile in referred patients with moderate chronic kidney disease: Composition of the German Chronic Kidney Disease (GCKD) cohort. *Nephrol. Dial. Transplant.* **30** 441–451.

TUBBS, J. D. (1986). Moments for a ratio of correlated gamma variates. *Comm. Statist. Theory Methods* **15** 251–259. MR0828616

TULUPYEV, A., SUVOROVA, A., SOUSA, J. and ZELTERMAN, D. (2013). Beta prime regression with application to risky behavior frequency screening. *Stat. Med.* **32** 4044–4056. MR3102433

WANG, T. and ZHAO, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.* **11** 771–791. MR3693546

WEGGEMANS, R. M. and TRAUTWEIN, E. A. (2003). Relation between soy-associated isoflavones and LDL and HDL cholesterol concentrations in humans: A meta-analysis. *Eur. J. Clin. Nutr.* **57** 940–946.

WEINHOLD, L., WAHL, S., PECHLIVANIS, S., HOFFMANN, P. and SCHMID, M. (2016). A statistical model for the analysis of beta values in DNA methylation studies. *BMC Bioinform*. **17** 480.

WIENS, B. L. (1999). When log-normal and gamma models give different results: A case study. *Amer. Statist*. **53** 89–93.

WILTFANG, J., ESSELMANN, H., BIBL, M., HÜLL, M., HAMPEL, H., KESSLER, H., FRÖLICH, L., SCHRÖDER, J., PETERS, O., JESSEN, F., LUCKHAUS, C., PERNECZKY, R., JAHN, H., FISZER, M., MALER, J. M., ZIMMERMANN, R., BRUCKMOSER, R., KORNHUBER, J. and LEWCZUK, P. (2007). Amyloid beta peptide ratio 42/40 but not A beta 42 correlates with phospho-Tau in patients with low- and high-CSF A beta 40 load. *J. Neurochem*. **101** 1053–1059.

YEE, T. W. (2015). *Vector Generalized Linear and Additive Models*: *With an Implementation in R*. Springer, New York. MR3408425

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 573–605
https://doi.org/10.1214/18-AOAS1200

DYNAMICS OF HOMELESSNESS IN URBAN AMERICA[1]

BY CHRIS GLYNN AND EMILY B. FOX

University of New Hampshire and University of Washington

The relationship between housing costs and homelessness has important implications for the way that city and county governments respond to increasing homeless populations. Though many analyses in the public policy literature have examined inter-community variation in homelessness *rates* to identify causal mechanisms of homelessness [*J. Urban Aff.* **35** (2013) 607–625; *J. Urban Aff.* **25** (2003) 335–356; *Am. J. Publ. Health* **103** (2013) S340–S347], few studies have examined time-varying homeless *counts* within the same community [*J. Mod. Appl. Stat. Methods* **15** (2016) 15]. To examine trends in homeless population counts in the 25 largest U.S. metropolitan areas, we develop a dynamic Bayesian hierarchical model for time-varying homeless count data. Particular care is given to modeling uncertainty in the homeless count generating and measurement processes, and a critical distinction is made between the counted number of homeless and the true size of the homeless population. For each metro under study, we investigate the relationship between increases in the Zillow Rent Index and increases in the homeless population. Sensitivity of inference to potential improvements in the accuracy of point-in-time counts is explored, and evidence is presented that the inferred increase in the rate of homelessness from 2011–2016 depends on prior beliefs about the accuracy of homeless counts. A main finding of the study is that the relationship between homelessness and rental costs is strongest in New York, Los Angeles, Washington, D.C., and Seattle.

REFERENCES

ALDOR-NOIMAN, S., BROWN, L. D., FOX, E. B. and STINE, R. A. (2016). Spatio-temporal low count processes with application to violent crime events. *Statist. Sinica* **26** 1587–1610. MR3586230

APPELBAUM, R. P., DOLNY, M., DREIER, P. and GILDERBLOOM, J. I. (1991). Scapegoating rent control: Masking the causes of homelessness. *J. Am. Plan. Assoc.* **57** 153–164.

BEEKMAN, D. (2016). King county's homeless count could soar with new method of tallying. *Seattle Times*. [Online; accessed 05/29/2017].

BEEKMAN, D. and BROOM, J. (2015). Mayor, county exec declare 'state of emergency' over homelessness. *Seattle Times*. [Online; accessed 04/15/2017].

BOHANON, C. (1991). The economic correlates of homelessness in sixty cities. *Soc. Sci. Q.*

BUN, Y. (2012). Zillow rent index: Methodology. https://www.zillow.com/research/zillow-rent-index-methodology-2393/ [Online; accessed 04/2/2017].

U.S. CENSUS BUREAU (2016). County population totals tables: 2010–2016. https://www.census.gov/data/tables/2016/demo/popest/counties-total.html [Online; accessed 04/2/2017].

BURT, M. M. (1992). Over the edge: The growth of homelessness in the 1980s. Russell Sage Foundation.

Key words and phrases. Homelessness, housing costs, missing data, state-space.

BYRNE, T., MUNLEY, E. A., FARGO, J. D., MONTGOMERY, A. E. and CULHANE, D. P. (2013). New perspectives on community-level determinants of homelessness. *J. Urban Aff.* **35** 607–625.
CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096
COLES, S. and SPARKS, R. (2006). Extreme value methods for modelling historical series of large volcanic magnitudes. *Statistics in Volcanology* **1** 47–56.
CORINTH, K. C. (2015). Ending homelessness: More housing or fewer shelters? AEI Economics Working Papers 863788.
CORNULIER, T., ROBINSON, R. A., ELSTON, D., LAMBIN, X., SUTHERLAND, W. J. and BENTON, T. G. (2011). Bayesian reconstitution of environmental change from disparate historical records: Hedgerow loss and farmland bird declines. *Methods Ecol. Evol.* **2** 86–94.
EARLY, D. W. and OLSEN, E. O. (2002). Subsidized housing, emergency shelters, and homelessness: An empirical investigation using data from the 1990 census. *Adv. Econ. Anal. Policy* **2**(1).
FARGO, J. D., MUNLEY, E. A., BYRNE, T. H., MONTGOMERY, A. E. and CULHANE, D. P. (2013). Community-level characteristics associated with variation in rates of homelessness among families and single adults. *Am. J. Publ. Health* **103**(S2) S340–S347.
FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889
GLYNN, C. and FOX, E. B. (2019). Supplement to "Dynamics of Homelessness in Urban America." DOI:10.1214/18-AOAS1200SUPP.
GRIMES, P. W. and CHRESSANTHIS, G. A. (1997). Assessing the effect of rent control on homelessness. *J. Urban Econ.* **41** 23–37.
HANRATTY, M. (2017). Do local economic conditions affect homelessness? Impact of area housing market factors, unemployment, and poverty on community homeless rates. *Hous. Policy Debate* 1–16.
HONIG, M. and FILER, R. K. (1993). Causes of intercity variation in homelessness. *Am. Econ. Rev.* 248–255.
HOPPER, K., SHINN, M., LASKA, E., MEISNER, M. and WANDERLING, J. (2008). Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *Am. J. Publ. Health* **98** 1438–1442.
HUDSON, C. G. (1998). Estimating homeless populations through structural equation modeling. *Soc. Choice Welf.* **25** 136.
KERY, M. and ROYLE, A. J. (2010). Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. *J. Anim. Ecol.* **79** 453–461.
LASKA, E. M. and MEISNER, M. (1993). A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 209–220.
LEE, B. A., PRICE-SPRATLEN, T. and KANAN, J. W. (2003). Determinants of homelessness in metropolitan areas. *J. Urban Aff.* **25** 335–356.
MCCANDLESS, L. C., PATTERSON, M. L., CURRIE, L. B., MONIRUZZAMAN, A. and SOMERS, J. M. (2016). Bayesian estimation of the size of a street-dwelling homeless population. *J. Mod. Appl. Stat. Methods* **15** 15.
O'FLAHERTY, B. (1995). An economic theory of homelessness and housing. *J. Hous. Econ.* **4** 13–49.
OFFICE OF COMMUNITY PLANNING AND DEVELOPMENT, DEPT. OF HOUSING AND URBAN DEVELOPMENT (2009). Continuum of care 101. https://www.hudexchange.info/resources/documents/CoC101.pdf. [Online; accessed 04/23/2018].
U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT (2016). Pit and hic data since 2007. https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/. [Online; accessed 04/2/2017].
POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

QUIGLEY, J. M. (1990). Does rent control cause homelessness? Taking the claim seriously. *J. Policy Anal. Manage.* **9** 89–93.

QUIGLEY, J. M. and RAPHAEL, S. (2001). The economics of homelessness: The evidence from North America. *European Journal of Housing Policy* **1** 323–336.

QUIGLEY, J. M., RAPHAEL, S. and SMOLENSKY, E. (2001). Homeless in America, homeless in California. *Rev. Econ. Stat.* **83** 37–51.

RAPHAEL, S. (2010). Housing market regulation and homelessness. In *How to House the Homeless* 110–135. Russell Sage Foundation, New York.

SCHWARZ, C. J. and SEBER, G. A. (1999). Estimating animal abundance: Review III. *Statist. Sci.* 427–456.

SCLAR, E. D. (1990). Homelessness and housing policy: A game of musical chairs. *Am. J. Publ. Health* **80** 1039–1040.

STOJANOVIC, D., WEITZMAN, B. C., SHINN, M., LABAY, L. E. and WILLIAMS, N. P. (1999). Tracing the path out of homelessness: The housing patterns of families after exiting shelter. *Am. J. Community Psychol.* **27** 199–208.

TOKDAR, S. T., GROSSMANN, I., KADANE, J. B., CHAREST, A.-S. and SMALL, M. J. (2011). Impact of beliefs about Atlantic tropical cyclone detection on conclusions about trends in tropical cyclone numbers. *Bayesian Anal.* **6** 547–572.

TROUTMAN, W., JACKSON, J. D. and EKELUND, R. B. (1999). Public policy, perverse incentives, and the homeless problem. *Public Choice* **98** 195–212.

WINDLE, J., POLSON, N. G. and SCOTT, J. G. (2014). Sampling Polya–Gamma random variates: Alternate and approximate techniques. Available at arXiv:1405.0506v1.

WINDLE, J., CARVALHO, C. M., SCOTT, J. G. and SUN, L. (2013). Efficient data augmentation in dynamic models for binary and count data. Available at arXiv:1308.0774.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 606–637
https://doi.org/10.1214/18-AOAS1208

BAYESIAN HIDDEN MARKOV TREE MODELS FOR CLUSTERING GENES WITH SHARED EVOLUTIONARY HISTORY[1]

BY YANG LI[*,§,2], SHAOYANG NING[*,2], SARAH E. CALVO[†,‡,§], VAMSI K. MOOTHA[¶,†,‡,§] AND JUN S. LIU[*]

Harvard University[*], Broad Institute[†], Harvard Medical School[‡], Massachusetts General Hospital[§], and Howard Hughes Medical Institute[¶]

Determination of functions for poorly characterized genes is crucial for understanding biological processes and studying human diseases. Functionally associated genes are often gained and lost together through evolution. Therefore identifying co-evolution of genes can predict functional gene-gene associations. We describe here the full statistical model and computational strategies underlying the original algorithm *CLustering by Inferred Models of Evolution* (CLIME 1.0) recently reported by us (*Cell* **158** (2014) 213–225). CLIME 1.0 employs a mixture of tree-structured hidden Markov models for gene evolution process, and a Bayesian model-based clustering algorithm to detect gene modules with shared evolutionary histories (termed evolutionary conserved modules, or ECMs). A Dirichlet process prior was adopted for estimating the number of gene clusters and a Gibbs sampler was developed for posterior sampling. We further developed an extended version, CLIME 1.1, to incorporate the uncertainty on the evolutionary tree structure. By simulation studies and benchmarks on real data sets, we show that CLIME 1.0 and CLIME 1.1 outperform traditional methods that use simple metrics (e.g., the Hamming distance or Pearson correlation) to measure co-evolution between pairs of genes.

REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—*1983. *Lecture Notes in Math*. **1117** 1–198. Springer, Berlin. MR0883646

BALSA, E., MARCO, R., PERALES-CLEMENTE, E., SZKLARCZYK, R., CALVO, E., LANDÁZURI, M. O. and ENRÍQUEZ, J. A. (2012). NDUFA4 is a subunit of complex IV of the mammalian electron transport chain. *Cell Metab.* **16** 378–386.

BARKER, D., MEADE, A. and PAGEL, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* **23** 14–20.

BARKER, D. and PAGEL, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput. Biol.* **1** e3.

BICK, A. G., CALVO, S. E. and MOOTHA, V. K. (2012). Evolutionary diversity of the mitochondrial calcium uniporter. *Science* **336** 886.

CHEN, R. and LIU, J. S. (1996). Predictive updating methods with application to Bayesian classification. *J. Roy. Statist. Soc. Ser. B* **58** 397–415. MR1377840

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

Key words and phrases. Co-evolution, Dirichlet process mixture model, evolutionary history, gene function prediction, tree-structured hidden Markov model.

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

GALPERIN, M. Y. and KOONIN, E. V. (2010). From complete genome sequence to "complete" understanding? *Trends Biotechnol.* **28** 398–406.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GLAZKO, G. V. and MUSHEGIAN, A. R. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol.* **5** R32.

GUINDON, S., DUFAYARD, J.-F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. and GASCUEL, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59** 307–321.

HAMMING, R. W. (1950). Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29** 147–160. MR0035935

HAMOSH, A., SCOTT, A. F., AMBERGER, J. S., BOCCHINI, C. A. and MCKUSICK, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33** D514–D517.

HORANI, A., FERKOL, T. W., DUTCHER, S. K. and BRODY, S. L. (2016). Genetics and biology of primary ciliary dyskinesia. *Paediatr. Respir. Rev.* **18** 18–24.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

INGLIS, P. N., BOROEVICH, K. A. and LEROUX, M. R. (2006). Piecing together a ciliome. *Trends Genet.* **22** 491–500.

JIM, K., PARMAR, K., SINGH, M. and TAVAZOIE, S. (2004). A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res.* **14** 109–115.

KENSCHE, P. R., VAN NOORT, V., DUTILH, B. E. and HUYNEN, M. A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. *J. R. Soc. Interface* **5** 151–170.

LI, J. B., GERDES, J. M., HAYCRAFT, C. J., FAN, Y., TESLOVICH, T. M., MAY-SIMERA, H., LI, H., BLACQUE, O. E., LI, L., LEITCH, C. C. et al. (2004). Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* **117** 541–552.

LI, Y., CALVO, S. E., GUTMAN, R., LIU, J. S. and MOOTHA, V. K. (2014). Expansion of biological pathways based on evolutionary inference. *Cell* **158** 213–225.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing. Springer Series in Statistics*. Springer, New York. MR2401592

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

MIMAKI, M., WANG, X., MCKENZIE, M., THORBURN, D. R. and RYAN, M. T. (2012). Understanding mitochondrial complex I assembly in health and disease. *Biochim. Biophys. Acta, Bioenerg.* **1817** 851–862.

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

OGILVIE, I., KENNAWAY, N. G. and SHOUBRIDGE, E. A. (2005). A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. *J. Clin. Invest.* **115** 2784–2792.

PAGEL, M. and MEADE, A. (2007). BayesTraits. Computer program and documentation. Available at http://www.evolution.rdg.ac.uk/bayestraits.html.

PAGLIARINI, D. J., CALVO, S. E., CHANG, B., SHETH, S. A., VAFAI, S. B., ONG, S. E., WALFORD, G. A. et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134** 112–123.

PELLEGRINI, M., MARCOTTE, E. M., THOMPSON, M. J., EISENBERG, D. and YEATES, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96** 4285–4288.

PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784

RONQUIST, F. and HUELSENBECK, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19** 1572–1574.

TABACH, Y., BILLI, A. C., HAYES, G. D., NEWMAN, M. A., ZUK, O., GABEL, H., KAMATH, R., YACOBY, K., CHAPMAN, B., GARCIA, S. M. et al. (2013). Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493** 694–698.

TRACHANA, K., LARSSON, T. A., POWELL, S., CHEN, W.-H., DOERKS, T., MULLER, J. and BORK, P. (2011). Orthology prediction methods: A quality assessment using curated protein families. *BioEssays* **33** 769–780.

VERT, J.-P. (2002). A tree kernel to analyse phylogenetic profiles. *Bioinformatics* **18** S276–S284.

VON MERING, C., HUYNEN, M., JAEGGI, D., SCHMIDT, S., BORK, P. and SNEL, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* **31** 258–261.

ZHOU, Y., WANG, R., LI, L., XIA, X. and SUN, Z. (2006). Inferring functional linkages between proteins from evolutionary scenarios. *J. Mol. Biol.* **359** 1150–1159.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 638–660
https://doi.org/10.1214/18-AOAS1209

SEQUENTIAL DIRICHLET PROCESS MIXTURES OF MULTIVARIATE SKEW *t*-DISTRIBUTIONS FOR MODEL-BASED CLUSTERING OF FLOW CYTOMETRY DATA[1]

BY BORIS P. HEJBLUM[*,†,2], CHARIFF ALKHASSIM[*,†], RAPHAEL GOTTARDO[‡], FRANÇOIS CARON[§] AND RODOLPHE THIÉBAUT[*,†]

Univ. Bordeaux[*], Vaccine Research Institute (VRI)[†], Fred Hutchinson Cancer Research Center[‡] and University of Oxford[§]

Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of a single cell. This enables us to identify cell subtypes and to determine their relative proportions. Improvements of this technology allow us to describe millions of individual cells from a blood sample using multiple markers. This results in high-dimensional datasets, whose manual analysis is highly time-consuming and poorly reproducible. While several methods have been developed to perform automatic recognition of cell populations most of them treat and analyze each sample independently. However, in practice individual samples are rarely independent, especially in longitudinal studies. Here we analyze new longitudinal flow-cytometry data from the DALIA-1 trial, which evaluates a therapeutic vaccine against HIV, by proposing a new Bayesian nonparametric approach with Dirichlet process mixture (DPM) of multivariate skew *t*-distributions to perform model based clustering of flow-cytometry data. DPM models directly estimate the number of cell populations from the data, avoiding model selection issues, and skew *t*-distributions provides robustness to outliers and nonelliptical shape of cell populations. To accommodate repeated measurements, we propose a sequential strategy relying on a parametric approximation of the posterior. We illustrate the good performance of our method on simulated data and on an experimental benchmark dataset. This sequential strategy outperforms all other methods evaluated on the benchmark dataset and leads to improved performance on the DALIA-1 data.

REFERENCES

AGHAEEPOUR, N., FINAK, G., HOOS, H., MOSMANN, T. R., BRINKMAN, R. R., GOTTARDO, R. and SCHEUERMANN, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques *Nat. Methods* **10** 228–238.

AGHAEEPOUR, N., NIKOLIC, R., HOOS, H. H. and BRINKMAN, R. R. (2011). Rapid cell population identification in flow cytometry data *Cytometry Part A* **79** 6–13.

AZZALINI, A., BROWNE, R. P., GENTON, M. G. and MCNICHOLAS, P. D. (2016). On nomenclature for, and the relative merits of, two formulations of skew distributions. *Statist. Probab. Lett.* **110** 201–206. MR3474758

Key words and phrases. Automatic gating, Bayesian nonparametrics, Dirichlet process, flow cytometry, HIV, mixture model, skew *t*-distribution.

AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 367–389. MR1983753

AZZALINI, A. and DALLA VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726. MR1440039

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38. MR0501592

BINDER, D. A. (1981). Approximations to Bayesian clustering rules. *Biometrika* **68** 275–285. MR0614964

BRINKMAN, R. R., GASPARETTO, M., LEE, S.-J. J., RIBICKAS, A. J., PERKINS, J., JANSSEN, W., SMILEY, R. and SMITH, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *J. Amer. Soc. Blood Marrow Transplantol. Biol. Blood Marrow Transplant.* **13** 691–700.

CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. and VANHEEGHE, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Trans. Signal Process.* **56** 71–84. MR2439814

CARON, F., NEISWANGER, W., WOOD, F., DOUCET, A. and DAVY, M. (2017). Generalized Pólya urn for time-varying Pitman–Yor processes. *J. Mach. Learn. Res.* **18** Paper No. 27. MR3634894

CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181. MR3262549

CHAN, C., FENG, F., OTTINGER, J., FOSTER, D., WEST, M. and KEPLER, T. B. (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry, Part A J. Internat. Soc. Anal. Cytol.* **73** 693–701.

CRON, A., GOUTTEFANGEAS, C., FRELINGER, J., LIN, L., SINGH, S. K., BRITTEN, C. M., WELTERS, M. J. P., VAN DER BURG, S. H., WEST, M. and CHAN, C. (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS Comput. Biol.* **9** e1003130.

DAHL, D. B. (2006). Model-based clustering for expression data via a dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* (K.-A. Do, P. Müller & M. Vannucci, eds.) 201–218. Cambridge Univ. Press, Cambridge. MR2269095

DUNDAR, M., AKOVA, F., YEREBAKAN, H. Z. and RAJWA, B. (2014). A non-parametric Bayesian model for joint cell clustering and cluster matching: Identification of anomalous sample phenotypes with random effects. *BMC Bioinform.* **15** 314.

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

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

FINAK, G., BASHASHATI, A., BRINKMAN, R. and GOTTARDO, R. (2009). Merging mixture components for cell population identification in flow cytometry. *Adv. Bioinform.* **2009** 247646.

FINAK, G., PEREZ, J.-M., WENG, A. and GOTTARDO, R. (2010). Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinform.* **11** 546.

FRITSCH, A. and ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4** 367–391. MR2507368

FRÜHWIRTH-SCHNATTER, S. and PYNE, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11** 317–336.

GE, Y. and SEALFON, S. C. (2012). flowPeaks: A fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* **28** 2052–2058.

GONDOIS-REY, F., GRANJEAUD, S., ROUILLIER, P., RIOUALEN, C., BIDAUT, G. and OLIVE, D. (2016). Multi-parametric cytometry from a complex cellular sample: Improvements and limits of manual versus computational-based interactive analyses. *Cytometry Part A* **89** 480–490.

HEJBLUM, B. P, ALKHASSIM, C., GOTTARDO, R., CARON, F. and THIÉBAUT, R. (2019). Supplement to "Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data." DOI:10.1214/18-AOAS1209SUPP.

HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8** 439–451. MR3066948

HUANG, Z. and GELMAN, A. (2005). Sampling for Bayesian computation with large datasets. *SSRN Electron. J.* 1–21.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

JOHNSSON, K., WALLIN, J. and FONTES, M. (2016). BayesFlow: Latent modeling of flow cytometry cell populations. *BMC Bioinform.* **17** 25.

JUÁREZ, M. A. and STEEL, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *J. Bus. Econom. Statist.* **28** 52–66. MR2650600

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

KESSLER, D. C., HOFF, P. D. and DUNSON, D. B. (2015). Marginally specified priors for nonparametric Bayesian estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 35–58. MR3299398

LARBI, A. and FULOP, T. (2014). From "truly naïve" to "exhausted senescent" T cells: When markers predict functionality. *Cytometry Part A* **85** 25–35.

LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. MR2351079

LEE, S. X. and MCLACHLAN, G. J. (2013). On mixtures of skew normal and skew t-distributions. *Adv. Data Anal. Classif.* **7** 241–266. MR3103965

LEE, S. X. and MCLACHLAN, G. J. (2016). Finite mixtures of canonical fundamental skew t-distributions. *Stat. Comput.* **26** 573–589. MR3489858

LÉVY, Y., THIÉBAUT, R., GOUGEON, M.-L., MOLINA, J.-M., WEISS, L., GIRARD, P.-M., VENET, A., MORLAT, P., POIRIER, B., LASCAUX, A.-S., BOUCHERIE, C., SERENI, D., ROUZIOUX, C., VIARD, J.-P., LANE, C., DELFRAISSY, J.-F., SERETI, I., CHÊNE, G. and ILIADE STUDY GROUP (2012). Effect of intermittent interleukin-2 therapy on CD4+ T-cell counts following antiretroviral cessation in patients with HIV. *AIDS* **26** 711–720.

LÉVY, Y., THIÉBAUT, R., MONTES, M., LACABARATZ, C., SLOAN, L., KING, B., PÉRUSAT, S., HARROD, C., COBB, A., ROBERTS, L. K., SURENAUD, M., BOUCHERIE, C., ZURAWSKI, S., DELAUGERRE, C., RICHERT, L., CHÊNE, G., BANCHEREAU, J. and PALUCKA, K. (2014). Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load. *Eur. J. Immunol.* **44** 2802–2810.

LIN, L., CHAN, C., HADRUP, S. R., FROESIG, T. M., WANG, Q. and WEST, M. (2013). Hierarchical Bayesian mixture modelling for antigen-specific T-cell subtyping in combinatorially encoded flow cytometry studies. *Stat. Appl. Genet. Mol. Biol.* **12** 309–331. MR3101032

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

LO, K., BRINKMAN, R. R. and GOTTARDO, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry, Part A J. Internat. Soc. Anal. Cytol.* **73** 321–332.

LO, K. and GOTTARDO, R. (2012). Flexible mixture modeling via the multivariate t distribution with the Box–Cox transformation: An alternative to the skew-t distribution. *Stat. Comput.* **22** 33–52. MR2865054

MCLACHLAN, G. J. and LEE, S. X. (2016). Comment on "On nomenclature, and the relative merits of two formulations of skew distributions" by A. Azzalini, R. Browne, M. Genton, and P. McNicholas. *Statist. Probab. Lett.* **116** 1–5. MR3508513

MEDVEDOVIC, M. and SIVAGANESAN, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18** 1194–1206.

MELCHIOTTI, R., GRACIO, F., KORDASTI, S., TODD, A. K. and DE RINALDIS, E. (2017). Cluster stability in the analysis of mass cytometry data. *Cytometry Part A* **91** 73–84.

MOSMANN, T. R., NAIM, I., REBHAHN, J., DATTA, S., CAVENAUGH, J. S., WEAVER, J. M. and SHARMA, G. (2014). SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation. *Cytometry Part A* **85** 422–433.

MURRAY, P. M., BROWNE, R. P. and MCNICHOLAS, P. D. (2014). Mixtures of skew-*t* factor analyzers. *Comput. Statist. Data Anal.* **77** 326–335. MR3210066

NAIM, I., DATTA, S., REBHAHN, J., CAVENAUGH, J. S., MOSMANN, T. R. and SHARMA, G. (2014). SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design. *Cytometry Part A* **85** 408–421.

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. MR1994729

PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002. With a foreword by Jean Picard. MR2245368

PYNE, S., HU, X., WANG, K., ROSSIN, E., LIN, T.-I., MAIER, L. M., BAECHER-ALLAN, C., MCLACHLAN, G. J., TAMAYO, P., HAFLER, D. A., DE JAGER, P. L. and MESIROV, J. P. (2009). Automated high-dimensional flow cytometric data analysis *Proc. Natl. Acad. Sci. USA* **106** 8519–8524.

QIAN, Y., WEI, C., EUN-HYUNG LEE, F., CAMPBELL, J., HALLILEY, J., LEE, J. A., CAI, J., KONG, Y. M., SADAT, E., THOMSON, E., DUNN, P., SEEGMILLER, A. C., KARANDIKAR, N. J., TIPTON, C. M., MOSMANN, T., SANZ, I. and SCHEUERMANN, R. H. (2010). Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry, Part B Clin. Cytom.* **78 Suppl 1** S69–82.

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

SUGÁR, I. P. and SEALFON, S. C. (2010). Misty Mountain clustering: Application to fast unsupervised flow cytometry gating. *BMC Bioinform.* **11** 502.

TEH, Y. W. (2010). Dirichlet process. In *Encyclopedia of Machine Learning* 280–287. Springer US, Boston, MA.

THIÉBAUT, R., PELLEGRIN, I., CHÊNE, G., VIALLARD, J. F., FLEURY, H., MOREAU, J. F., PELLEGRIN, J. L. and BLANCO, P. (2005). Immunological markers after long-term treatment interruption in chronically HIV-1 infected patients with CD4 cell count above 400 × 10(6) cells/l. *AIDS* **19** 53–61.

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

VAN DYK, D. A. and JIAO, X. (2015). Metropolis–Hastings within partially collapsed Gibbs samplers. *J. Comput. Graph. Statist.* **24** 301–327. MR3357383

VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103** 790–796. MR2524010

WELTERS, M. J. P., GOUTTEFANGEAS, C., RAMWADHDOEBE, T. H., LETSCH, A., OTTENSMEIER, C. H., BRITTEN, C. M. and VAN DER BURG, S. H. (2012). Harmonization of the intracellular cytokine staining assay. *Cancer Immunol. Immunother.* **61** 967–978.

ZARE, H., SHOOSHTARI, P., GUPTA, A. and BRINKMAN, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinform.* **11** 403.

The Annals of Applied Statistics
2019, Vol. 13, No. 1, 661–681
https://doi.org/10.1214/18-AOAS1210

COMPOSITIONAL MEDIATION ANALYSIS FOR MICROBIOME STUDIES[1]

BY MICHAEL B. SOHN AND HONGZHE LI[2]

University of Pennsylvania

Motivated by recent advances in causal mediation analysis and problems in the analysis of microbiome data, we consider the setting where the effect of a treatment on an outcome is transmitted through perturbing the microbial communities or compositional mediators. The compositional and high-dimensional nature of such mediators makes the standard mediation analysis not directly applicable to our setting. We propose a sparse compositional mediation model that can be used to estimate the causal direct and indirect (or mediation) effects utilizing the algebra for compositional data in the simplex space. We also propose tests of total and component-wise mediation effects. We conduct extensive simulation studies to assess the performance of the proposed method and apply the method to a real microbiome dataset to investigate an effect of fat intake on body mass index mediated through the gut microbiome.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. CRC Press, London. MR0865647

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

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182.

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BILLHEIMER, D., GUTTORP, P. and FAGAN, W. F. (2001). Statistical interpretation of species composition. *J. Amer. Statist. Assoc.* **96** 1205–1214. MR1946574

BOLLEN, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociol. Method.* **17** 37–69.

BRAY, G. A. and POPKIN, B. M. (1998). Dietary fat intake does affect obesity! *Am. J. Clin. Nutr.* **68** 1157–1173.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CHÉN, O. Y., CRAINICEANU, C., OGBURN, E. L., CAFFO, B. S., WAGER, T. D. and LINDQUIST, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19** 121–136. MR3799607

Key words and phrases. Compositional algebra, 16S sequencing, causal mediation effect, simplex space.

DANIEL, H., GHOLAMI, A. M., BERRY, D., DESMARCHELIER, C., HAHNE, H., LOH, G., MONDOT, S., LEPAGE, P., ROTHBALLER, M., WALKER, A., BÖHM, C., WENNING, M., WAGNER, M., BLAUT, M., SCHMITT-KOPPLIN, P., KUSTER, B., HALLER, D. and CLAVEL, T. (2014). High-fat diet alters gut microbiota physiology in mice. *ISEM J.* **8** 295–308.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903

HU, F. B., RIMM, E., SMITH-WARNER, S. A., FESKANICH, D., STAMPFER, M. J., ASCHERIO, A., SAMPSON, L. and WILLETT, W. C. (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am. J. Clin. Nutr.* **69** 243–249.

HUANG, Y.-T. and PAN, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72** 401–413. MR3515767

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334.

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. MR2741814

IMAI, K. and YAMAMOTO, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Polit. Anal.* **21** 141–171.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—For Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951

LAM, Y. Y., HA, C. W. Y., CAMPBELL, C. R., MITCHELL, A. J., DINUDOM, A., OSCARSSON, J., COOK, D. I., HUNT, N. H., CATERSON, I. D., HOLMES, A. J. and STORLIEN, L. H. (2012). Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. *PLoS ONE* **7** e34233.

LEY, R. E., TURNBAUGH, P. J., KLEIN, S. and GORDON, J. I. (2006). Human gut microbes associated with obesity. *Nature* **444** 1022–1023.

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

MACHADO, J. A. F. and PARENTE, P. (2005). Bootstrap estimation of covariance matrices via the percentile method. *Econom. J.* **8** 70–78. MR2136930

MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. and SHEETS, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* **7** 83–104.

MAIER, M. J. (2014). DirichletReg: Dirichlet regression for compositional data in R. Research Report Series, Dept. Statistics and Mathematics, 125. WU Vienna Univ. Economics and Business, Vienna.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* 411–420. Morgan Kaufmann, San Francisco, CA.

PREACHER, K. J. and HAYES, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* **40** 879–891.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

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

SHROUT, P. E. and BOLGER, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychol. Methods* **7** 422–445.

SOBEL, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Method.* **13** 290–312.

SOHN, M. B. and LI, H. (2019). Supplement to "Compositional mediation analysis for microbiome studies." DOI:10.1214/18-AOAS1210SUPP.

TEIXEIRA, T. F., COLLADO, M. C., FERREIRA, C. L., BRESSAN, J. and PELUZIO, M. C. (2012). Potential mechanisms for the emerging link between obesity and increased intestinal permeability. *Nutr. Res*. **32** 637–47.

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.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol*. **172** 1339–1348.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiol. Methods* **2** 95–115.

WINSHIP, C. and MARE, R. D. (1983). Structural equations and path analysis for discrete data. *Amer. J. Sociol*. **89** 54–110.

WU, G., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y. Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R., SINHA, R., GILROY, E., GUPTA, K., BALDASSANO, R., NESSEL, L., LI, H., BUSHMAN, F. D. and LEWIS, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

ZHAO, Y. and LUO, X. (2016). Pathway lasso: Estimate and select sparse mediation pathways with high dimensional mediators. arXiv:1603.07749.

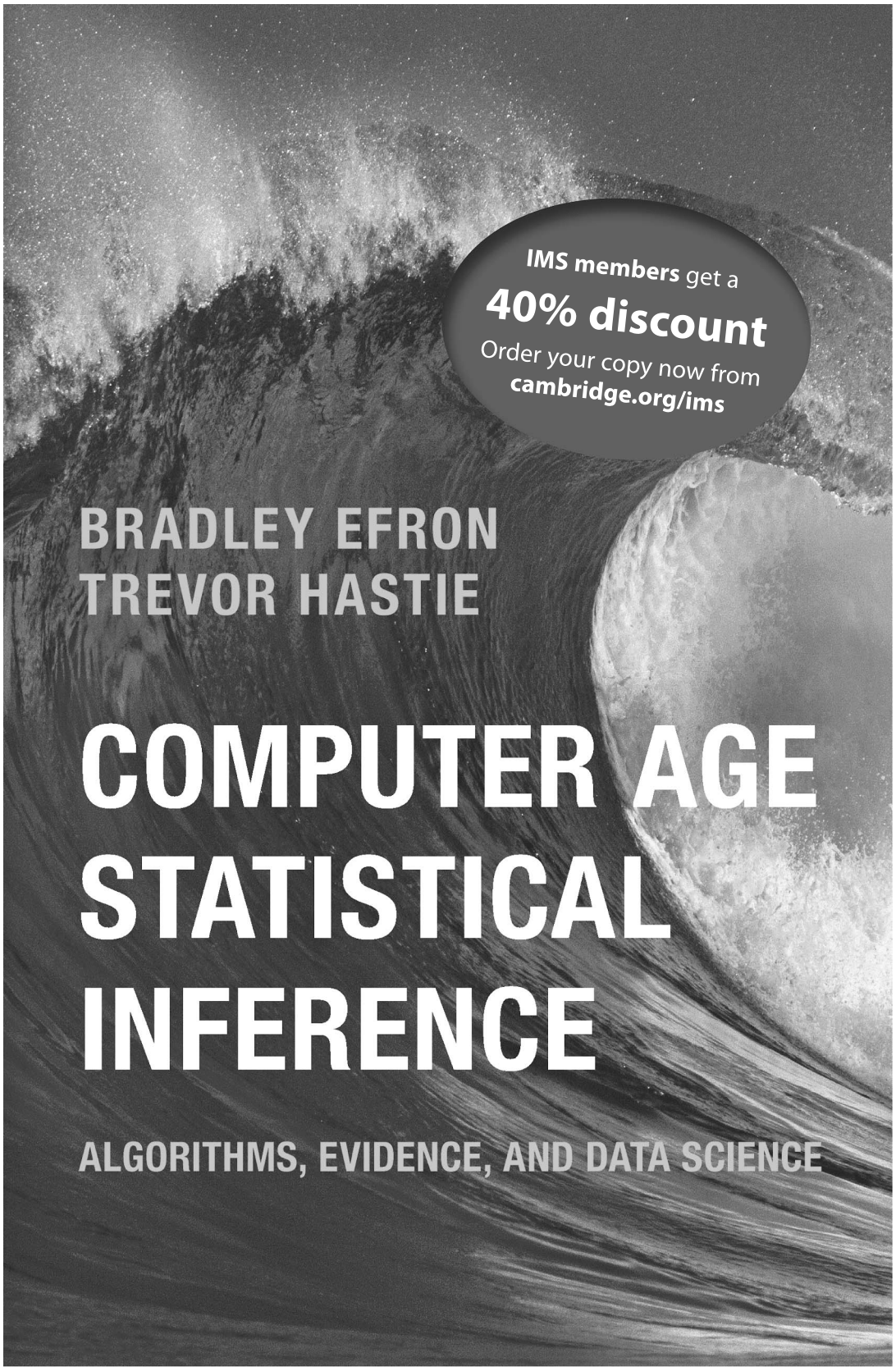
The Annals of Applied Statistics

Next Issues

Continued

The Annals of Applied Statistics

Next Issues–*Continued*


IMS members get a
40% discount
Order your copy now from
cambridge.org/ims
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE
STATISTICAL
INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE