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. 15, No. 3, pp. 1081–1581 September 2021

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: Krzysztof Burdzy, Department of Mathematics, University of Washington, Seattle, Washington 98195-4350, USA

President-Elect: Peter Bühlmann, Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland

Past President: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

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

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

Program Secretary: Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, 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, Charlottesville, VA 22904-4135, USA

The Annals of Probability. *Editors*: Alice Guionnet, Unité de Mathématiques Pures et Appliquées, ENS de Lyon, Lyon, France. Christophe Garban, Institut Camille Jordan, Université Claude Bernard Lyon 1, 69622 Villeurbanne, France

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*: Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

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 15, Number 3, September 2021. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, 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, PO Box 729, Middletown, Maryland 21769, USA.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1081–1083
https://doi.org/10.1214/21-AOAS1500

EDITORIAL: STATISTICAL SIGNIFICANCE, *P*-VALUES, AND REPLICABILITY

BY KAREN KAFADAR

Editor-in-Chief, 2019–2021

kk3ab@virginia.edu

REFERENCES

AMRHEIN, V., GREENLAND, S. and MCSHANE, B. (2019). Scientists rise up against statistical significance. *Nature* **20** 305–307.

BEIGEL, J. H., TOMASHEK, K. M., DODD, L. E. et al. (2020). Remdesivir for the treatment of Covid-19—final report. *N. Engl. J. Med.* **383** 1813–1826. https://doi.org/10.1056/NEJMoa200776

COX, D. R. (1986). Some general aspects of the theory of statistics. *International Statistical Review* **54** 117–126.

COX, D. R. (2020). Discussion of paper by Brad Efron. *Journal of the American Statistical Association* **115** 659–659.

DENWORTH, L. (2019). A Significant Problem. *Sci. Am.* **10** 63–67. 2019.

FEDERAL JUDICIAL CENTER (2011). *Reference Manual on Scientific Evidence*, 3rd ed., National Academies Press, Washington, DC.

FISHER, R. (1935). *The Design of Experiments*. Oliver and Boyd, London.

KAYE, D. H. and FREEDMAN, D. A. (2011). Reference guide on statistics. In *Reference Manual on Scientific Evidence* Third Edition 211–302, National Academies Press.

MATRIXX INITIATIVES, INC. V. SIRACUSANO, 563 U.S. 27 (2011). 131 S.Ct. 1309 Supreme Court of the United States, No. 09-1156 (179 L.Ed.2d 398, 79 USLW 4187, Fed. Sec. L. Rep. P 96,249.)

REID, N. and COX, D. R. (2014). On some principles of statistical inference. *Int. Stat. Rev.* **83** 293–308. https://doi.org/10.1111/insr.12067

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856 https://doi.org/10.1073/pnas.1530509100

TARRAN, B. (2019). The *S* word… and what to do about it. *Signficance*, August 2019: 14. [Correction to initial version published 24 July 2019: "Correction added on 26 February 2021, after publication: This article has been updated to clarify that the recommendation to abandon the term 'statistical significance' came from the editors of a special issue of The American Statistician. not the American Statistical Association (which publishes the journal). We apologise for any confusion."]

TUKEY, J. W. (1969). Analyzing data: Sanctification or detective work. *Amer. Psychol.* **24** 83–91.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1084–1085
https://doi.org/10.1214/21-AOAS1501

THE ASA PRESIDENT'S TASK FORCE STATEMENT ON STATISTICAL SIGNIFICANCE AND REPLICABILITY

BY YOAV BENJAMINI[1], RICHARD D. DE VEAUX[2], BRADLEY EFRON[3], SCOTT EVANS[4], MARK GLICKMAN[5,*], BARRY I. GRAUBARD[6], XUMING HE[7], XIAO-LI MENG[5,†], NANCY REID[8], STEPHEN M. STIGLER[9], STEPHEN B. VARDEMAN[10], CHRISTOPHER K. WIKLE[11], TOMMY WRIGHT[12], LINDA J. YOUNG[13] AND KAREN KAFADAR[14]

[1]*Department of Statistics and Operations Research, Tel Aviv University, ybenja@tauex.tau.ac.il*

[2]*Department of Mathematics and Statistics, Williams College, deveaux@williams.edu*

[3]*Department of Statistics and Department of Biomedical Data Sciences, Stanford University, brad@stat.stanford.edu*

[4]*Department of Biostatistics & Bioinformatics, George Washington University, sevans@bsc.gwu.edu*

[5]*Department of Statistics, Harvard University, *glickman@fas.harvard.edu; †meng@stat.harvard.edu*

[6]*Biostatistics Branch, National Cancer Institite, barry.graubard@nih.gov*

[7]*(Co-chair), Department of Statistics, University of Michigan, xmhe@umich.edu*

[8]*Department of Statistics, University of Toronto, reid@utstat.utoronto.ca*

[9]*Department of Statistics, University of Chicago, stigler@uchicago.edu*

[10]*Department of Statistics and Department of Industrial & Manufacturing Systems Engineering, Iowa State University, vardeman@iastate.edu*

[11]*Department of Statistics, University of Missouri, wiklec@missouri.edu*

[12]*Center for Statistical Research and Methodology, United States Bureau of the Census, tommy.wright@census.gov*

[13]*(Co-chair), Research & Development, National Agricultural Statistics Service, linda.j.young@usda.gov*

[14]*(Ex-officio), Department of Statistics, University of Virginia, kkafadar@virginia.edu*

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1086–1100
https://doi.org/10.1214/21-AOAS1481

A DEEP LEARNING SEMIPARAMETRIC REGRESSION FOR ADJUSTING COMPLEX CONFOUNDING STRUCTURES

BY XINLEI MI[1], PATRICK TIGHE[2], FEI ZOU[3,*] AND BAIMING ZOU[3,†]

[1]*Department of Preventive Medicine Biostatistics, Northwestern University, xinlei.mi@northwestern.edu*
[2]*Department of Anesthesiology, University of Florida, ptighe@anest.ufl.edu*
[3]*Department of Biostatistics, University of North Carolina at Chapel Hill, *feizou@email.unc.edu; †bzou@email.unc.edu*

Deep Treatment Learning (deepTL), a robust yet efficient deep learning-based semiparametric regression approach, is proposed to adjust the complex confounding structures in comparative effectiveness analysis of observational data, for example, electronic health record (EHR) data in which complex confounding structures are often embedded. Specifically, we develop a deep learning neural network with a score-based ensembling scheme for flexible function approximation. An improved semiparametric procedure is further developed to enhance the performance of the proposed method under finite sample settings. Comprehensive numerical studies have demonstrated the superior performance of the proposed methods, as compared with existing methods, with a remarkably reduced bias and mean squared error in parameter estimates. The proposed research is motivated by a postsurgery pain study, which is also used to illustrate the practical application of deepTL. Finally, an R package, "deepTL," is developed to implement the proposed method.

REFERENCES

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424. https://doi.org/10.1080/00273171.2011.568786

BENGIO, Y., DELALLEAU, O. and ROUX, N. L. (2006). The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems* 107–114.

BENSON, K. and HARTZ, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med.* **342** 1878–1886.

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BRITTON, A., MCKEE, M., BLACK, N., MCPHERSON, K., SANDERSON, C. and BAIN, C. (1997). Choosing between randomised and non-randomised studies: A systematic review. *Health Technology Assessment* (*Winchester, England*) **2** i–iv.

BYRD, R. H., CHIN, G. M., NOCEDAL, J. and WU, Y. (2012). Sample size selection in optimization methods for machine learning. *Math. Program.* **134** 127–155. MR2947555 https://doi.org/10.1007/s10107-012-0572-5

CHEN, Y., CARROLL, R. J., HINZ, E. R. M., SHAH, A., EYLER, A. E., DENNY, J. C. and XU, H. (2013). Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *J. Am. Med. Inform. Assoc.* **20** e253–e259.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. K. (2016). Double machine learning for treatment and causal parameters. Technical Report No. CWP49/16. CeMMAP working paper, Centre for Microdata Methods and Practice, London.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CURRY, J. I., REEVES, B. and STRINGER, M. D. (2003). Randomized controlled trials in pediatric surgery: Could we do better? *Journal of Pediatric Surgery* **38** 556–559.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

ENGLE, R. F., GRANGER, C. W., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

Key words and phrases. Bootstrap aggregating, comparative effectiveness analysis, complex confounding, deep neural network, propensity score, semiparametric regression.

FARAGÓ, A. and LUGOSI, G. (1993). Strong universal consistency of neural network classifiers. *IEEE Trans. Inf. Theory* **39** 1146–1151.
HANSEN, L. K. and SALAMON, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 993–1001.
HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** 359–366.
KAZLEY, A. S. and OZCAN, Y. A. (2008). Do hospitals with electronic medical records (EMRs) provide higher quality care?: An examination of three clinical conditions. *Med. Care Res. Rev.* **65** 496–513. https://doi.org/10.1177/1077558707313437
KINGA, D. and ADAM, J. B. (2015). A method for stochastic optimization. In *International Conference on Learning Representations* (*ICLR*) **5**.
LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.
MACLEHOSE, R., REEVES, B., HARVEY, I., SHELDON, T., RUSSELL, I. and BLACK, A. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* (*Winchester, England*) **4** 1–154.
MCCULLOCH, P., TAYLOR, I., SASAKO, M., LOVETT, B. and GRIFFIN, D. (2002). Randomised trials in surgery: Problems and possible solutions. *BMJ, Br. Med. J.* **324** 1448–1451.
MI, X., TIGHE, P., ZOU, F. and ZOU, B. (2021). Supplement to "A deep learning semiparametric regression for adjusting complex confounding structures." https://doi.org/10.1214/21-AOAS1481SUPPA, https://doi.org/10.1214/21-AOAS1481SUPPB
MI, X., ZOU, F. and ZHU, R. (2019). Bagging and deep learning in optimal individualized treatment rules. *Biometrics* **75** 674–684. MR3999189 https://doi.org/10.1111/biom.12990
MIRIOVSKY, B. J., SHULMAN, L. N. and ABERNETHY, A. P. (2012). Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. *J. Clin. Oncol.* **30** 4243–4248.
MURDOCH, T. B. and DETSKY, A. S. (2013). The inevitable application of big data to health care. *JAMA* **309** 1351–1352.
NIE, X. and WAGER, S. (2017). Learning objectives for treatment effect estimation. Preprint. Available at arXiv:1712.04912.
PLATT, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* **10** 61–74.
PSATY, B. M. and LARSON, E. B. (2013). Investments in infrastructure for diverse research resources and the health of the public. *JAMA* **309** 1895–1896.
ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730
ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705
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 https://doi.org/10.1093/biomet/70.1.41
SCHISTERMAN, E. F., COLE, S. R. and PLATT, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* (*Cambridge, Mass.*) **20** 488–495.
SHAH, N. H. and TENENBAUM, J. D. (2012). Focus on translational bioinformatics: The coming age of data-driven medicine: Translational bioinformatics' next frontier. *J. Am. Med. Inform. Assoc.* **19** e2.
SHIR, Y., RAJA, S. N. and FRANK, S. M. (1994). The effect of epidural versus general anesthesia on postoperative pain and analgesic requirements in patients undergoing radical prostatectomy. *Anesthesiology* **80** 49–56.
SONODA, S. and MURATA, N. (2017). Neural network with unbounded activation functions is universal approximator. *Appl. Comput. Harmon. Anal.* **43** 233–268. MR3668038 https://doi.org/10.1016/j.acha.2015.12.005
STOCK, J. H. (1991). Nonparametric policy analysis: An application to estimating hazardous waste cleanup benefits. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* 77–98. Cambridge University Press, Cambridge.
TIGHE, P. J., KING, C. D., ZOU, B. and FILLINGIM, R. B. (2016). Time to onset of sustained postoperative pain relief (SuPPR): Evaluation of a new systems-level metric for acute pain management. *The Clinical Journal of Pain* **32** 371–379.
TIPPING, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1** 211–244. MR1875838 https://doi.org/10.1162/15324430152748236
TOH, S., GARCÍA RODRÍGUEZ, L. A. and HERNÁN, M. A. (2011). Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: An application to electronic medical records. *Pharmacoepidemiol. Drug Saf.* **20** 849–857.
TU, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49** 1225–1231.

TVERSKOY, M., COZACOV, C., AYACHE, M., BRADLEY, J. E. and KISSIN, I. (1990). Postoperative pain after inguinal herniorrhaphy with different types of anesthesia. *Anesthesia and Analgesia* **70** 29–35.

WILLIAMS, C. K. and BARBER, D. (1998). Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 1342–1351.

ZHOU, Z.-H., WU, J. and TANG, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence* **137** 239–263. MR1906477 https://doi.org/10.1016/S0004-3702(02)00190-X

ZOU, B., ZOU, F., SHUSTER, J. J., TIGHE, P. J., KOCH, G. G. and ZHOU, H. (2016). On variance estimate for covariate adjustment by propensity score analysis. *Stat. Med.* **35** 3537–3548. MR3537221 https://doi.org/10.1002/sim.6943

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1101–1125
https://doi.org/10.1214/20-AOAS1424

UNSUPERVISED STREAMING ANOMALY DETECTION FOR INSTRUMENTED INFRASTRUCTURE

BY HENRIQUE HOELTGEBAUM[*], NIALL ADAMS[†] AND F. DIN-HOUN LAU[‡]

Department of Mathematics, Imperial College London, [*]hh3015@imperial.ac.uk; [†]n.adams@imperial.ac.uk; [‡]dhlau55@gmail.com

Structural health monitoring (SHM) often involves instrumenting structures with distributed sensor networks. These networks typically provide high frequency data describing the spatiotemporal behaviour of the assets. A main objective of SHM is to reason about changes in structures' behaviour using sensor data. We construct a streaming anomaly detection method for data from a railway bridge instrumented with a fibre-optic sensor network. The data exhibits trend over time, which may be partially attributable to environmental factors, calling for temporally adaptive estimation. Exploiting a latent structure present in the data motivates a quantity of interest for anomaly detection. This quantity is estimated, sequentially and adaptively, using a new formulation of streaming principal component analysis. Anomaly detection for this quantity is then provided using conformal prediction. Like all streaming methods, the proposed method has free control parameters which are set using simulations based on bridge data. Experiments demonstrate that this method can operate at the sampling frequency of the data while providing accurate tracking of the target quantity. Further, the anomaly detection is able to detect train passage events. Finally, the method reveals a previously unreported cyclic structure present in the data.

REFERENCES

AGGARWAL, C. C. (2007). *Data Streams*: *Models and Algorithms* **31**. Springer, Berlin.

ANAGNOSTOPOULOS, C., TASOULIS, D. K., ADAMS, N. M., PAVLIDIS, N. G. and HAND, D. J. (2012). Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Stat. Anal. Data Min.* **5** 139–166. MR2910024 https://doi.org/10.1002/sam.10151

BALASUBRAMANIAN, V., HO, S.-S. and VOVK, V. (2014). *Conformal Prediction for Reliable Machine Learning*: *Theory*, *Adaptations and Applications*. Newnes, London.

BALZANO, L., CHI, Y. and LU, Y. M. (2018). Streaming PCA and subspace tracking: The missing data case. *Proc. IEEE* **106** 1293–1310.

BENCZÚR, A. A., KOCSIS, L. and PÁLOVICS, R. (2018). Online machine learning in big data streams. Preprint. Available at arXiv:1802.05872.

BODENHAM, D. A. and ADAMS, N. M. (2017). Continuous monitoring for changepoints in data streams using adaptive estimation. *Stat. Comput.* **27** 1257–1270. MR3647096 https://doi.org/10.1007/s11222-016-9684-8

BOUTSIDIS, C., GARBER, D., KARNIN, Z. and LIBERTY, E. (2015). Online principal components analysis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* 887–901. SIAM, Philadelphia, PA. MR3451085 https://doi.org/10.1137/1.9781611973730.61

BOWERS, K., BUSCHER, V., DENTTEN, R., EDWARDS, M., ENGLAND, J., ENZER, M., PARLIKAD, A. K. and SCHOOLING, J. (2016). Smart infrastructure: Getting more from strategic assets. Centre for Smart Infrastructure and Construction.

BURNAEV, E. and VOVK, V. (2014). Efficiency of conformalized ridge regression. In *Conference on Learning Theory* 605–622.

BUTLER, L. J., GIBBONS, N., HE, P., MIDDLETON, C. and ELSHAFIE, M. Z. (2016a). Evaluating the early-age behaviour of full-scale prestressed concrete beams using distributed and discrete fibre optic sensors. *Construction and Building Materials* **126** 894–912.

Key words and phrases. Structural health monitoring, streaming PCA, stochastic gradient descent, adaptive estimation, conformal prediction.

BUTLER, L. J., GIBBONS, N., HE, P., MIDDLETON, C. and ELSHAFIE, M. Z. (2016b). Evaluating the early-age behaviour of full-scale prestressed concrete beams using distributed and discrete fibre optic sensors. *Construction and Building Materials* **126 (Supplement C)** 894–912.

BUTLER, L. J., XU, J., HE, P., GIBBONS, N., DIRAR, S., MIDDLETON, C. R. and ELSHAFIE, M. Z. (2018). Robust fibre optic sensor arrays for monitoring early-age performance of mass-produced concrete sleepers. *Structural Health Monitoring* **17** 635–653.

CARDOT, H. and DEGRAS, D. (2018). Online principal component analysis in high dimension: Which algorithm to choose? *Int. Stat. Rev.* **86** 29–50. MR3796510 https://doi.org/10.1111/insr.12220

CHAMP, C. W. and WOODALL, W. H. (1987). Exact results for Shewhart control charts with supplementary runs rules *Technometrics* **29** 393–399.

CHERNOZHUKOV, V., WUTHRICH, K. and ZHU, Y. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. Preprint. Available at arXiv:1802.06300.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836. MR0556476

DAS, S., SAHA, P. and PATRO, S. (2016). Vibration-based damage detection techniques used for health monitoring of structures: A review. *Journal of Civil Structural Health Monitoring* **6** 477–507.

DOMINGOS, P. and HULTEN, G. (2003). A general framework for mining massive data streams. *J. Comput. Graph. Statist.* **12** 945–949.

FARRAR, C. R. and WORDEN, K. (2006). An introduction to structural health monitoring. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 303–315.

GAMA, J. (2010). *Knowledge Discovery from Data Streams*. *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*. CRC Press, Boca Raton, FL. MR2779331

GIRAUD, L., LANGOU, J. and ROZLOZNIK, M. (2005). The loss of orthogonality in the Gram–Schmidt orthogonalization process. *Comput. Math. Appl.* **50** 1069–1075. MR2167744 https://doi.org/10.1016/j.camwa.2005.08.009

GLISIC, B., INAUDI, D., LAU, J. M., MOK, Y. C. and NG, C. T. (2005). Long-term monitoring of high-rise buildings using long-gauge fibre optic sensors. In 7*th International Conference on Multi-Purpose High-Rise Towers and Tall Buildings*, *Dubai*, *UAM*, 10–11 *December* (*on Conference CD*, *Paper* #0416).

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

HAYKIN, S. S. (2008). *Adaptive Filter Theory*. Pearson, Upper Saddle River.

HERNANDEZ-GARCIA, M. R. and MASRI, S. F. (2014). Application of statistical monitoring using latent-variable techniques for detection of faults in sensor networks. *Journal of Intelligent Material Systems and Structures* **25** 121–136.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

JOLLIFFE, I. T. (2011). *Principal Component Analysis*. Springer.

KIM, A. Y., MARZBAN, C., PERCIVAL, D. B. and STUETZLE, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Process.* **89** 2529–2536.

LAU, F. D.-H., ADAMS, N. M., GIROLAMI, M. A., BUTLER, L. J. and ELSHAFIE, M. Z. E. B. (2018a). The role of statistics in data-centric engineering. *Statist. Probab. Lett.* **136** 58–62. MR3806838 https://doi.org/10.1016/j.spl.2018.02.035

LAU, F. D. -H., BUTLER, L. J., ADAMS, N. M., ELSHAFIE, M. Z. E. B. and GIROLAMI, M. A. (2018b). Real-time statistical modelling of data generated from self-sensing bridges. *Proceedings of the Institution of Civil Engineers—Smart Infrastructure and Construction* **171** 3–13.

LAXHAMMAR, R. and FALKMAN, G. (2015). Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Ann. Math. Artif. Intell.* **74** 67–94. MR3353897 https://doi.org/10.1007/s10472-013-9381-7

LEI, J., RINALDO, A. and WASSERMAN, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 71–96. MR3153934 https://doi.org/10.1111/rssb.12021

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

MEASURES, R. M., LEBLANC, M., LIU, K., FERGUSON, S., VALIS, T., HOGG, D., TURNER, R. and MCEWEN, K. (1992). Fiber optic sensors for smart structures. *Optics and Lasers in Engineering* **16** 127–152.

MEZZADRI, F. (2007). How to generate random matrices from the classical compact groups. *Notices Amer. Math. Soc.* **54** 592–604. MR2311982

MICRON OPTICS (2013). ENLIGHT User Guide. Available at http://www.micronoptics.com/download/enlight-user-guide-revision-1-138/#. Accessed: 2019-04-06.

MITLIAGKAS, I., CARAMANIS, C. and JAIN, P. (2013). Memory limited, streaming PCA. In *Advances in Neural Information Processing Systems* 2886–2894.

NADLER, B. (2011). On the distribution of the ratio of the largest eigenvalue to the trace of a Wishart matrix. *J. Multivariate Anal.* **102** 363–371. MR2739121 https://doi.org/10.1016/j.jmva.2010.10.005

NOVEMBRE, J. and STEPHENS, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40** 646–649.

OJA, E. (1992). Principal components, minor components, and linear neural networks. *Neural Netw.* **5** 927–935.

OJA, E. and KARHUNEN, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *J. Math. Anal. Appl.* **106** 69–84. MR0780319 https://doi.org/10.1016/0022-247X(85)90131-3

SANGER, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Netw.* **2** 459–473.

SCHOLZ, M. (2007). Analysing periodic phenomena by circular PCA. In *International Conference on Bioinformatics Research and Development* 38–47. Springer, Berlin.

STEWART, G. W. (1980). The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numer. Anal.* **17** 403–409. MR0581487 https://doi.org/10.1137/0717034

TODD, M. D., NICHOLS, J. M., TRICKEY, S. T., SEAVER, M., NICHOLS, C. J. and VIRGIN, L. N. (2006). Bragg grating-based fibre optic sensors in structural health monitoring. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 317–343.

VOVK, V. (2013). Conditional validity of inductive conformal predictors. *Mach. Learn.* **92** 349–376. MR3080332 https://doi.org/10.1007/s10994-013-5355-6

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Conformal Prediction*. Springer, Berlin.

VOVK, V., NOURETDINOV, I. and GAMMERMAN, A. (2009). On-line predictive linear regression. *Ann. Statist.* **37** 1566–1590. MR2509084 https://doi.org/10.1214/08-AOS622

WARMUTH, M. K. and KUZMIN, D. (2008). Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *J. Mach. Learn. Res.* **9** 2287–2320. MR2452628

WENG, J., ZHANG, Y. and HWANG, W.-S. (2003). Candid covariance-free incremental principal component analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 1034–1040.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1126–1146
https://doi.org/10.1214/21-AOAS1482

EXTENDING MODELS VIA GRADIENT BOOSTING: AN APPLICATION TO MENDELIAN MODELS

BY THEODORE HUANG[1,2,*], GREGORY IDOS[3,§], CHRISTINE HONG[3,¶], STEPHEN B. GRUBER[3,‖], GIOVANNI PARMIGIANI[1,2,†] AND DANIELLE BRAUN[1,2,‡]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [*]*thuang@ds.dfci.harvard.edu;* [†]*gp@jimmy.harvard.edu;* [‡]*dbraun@mail.harvard.edu*

[2]*Department of Data Science, Dana-Farber Cancer Institute*

[3]*City of Hope Comprehensive Cancer Center,* [§]*gidos@coh.org;* [¶]*chhong@coh.org;* [‖]*sgruber@coh.org*

Improving existing widely-adopted prediction models is often a more efficient and robust way toward progress than training new models from scratch. Existing models may: (a) incorporate complex mechanistic knowledge, (b) leverage proprietary information, and (c) have surmounted barriers to adoption. Compared to model training, model improvement and modification receive little attention. In this paper we propose a general approach to model improvement: we combine gradient boosting with any previously developed model to improve model performance while retaining important existing characteristics. To exemplify, we consider the context of Mendelian models which estimate the probability of carrying genetic mutations that confer susceptibility to disease by using family pedigrees and health histories of family members. Via simulations, we show that integration of gradient boosting with an existing Mendelian model can produce an improved model that outperforms both that model and the model built using gradient boosting alone. We illustrate the approach on genetic testing data from the USC–Stanford Cancer Genetics Hereditary Cancer Panel (HCP) study.

REFERENCES

ANTONIOU, A., CUNNINGHAM, A., PETO, J., EVANS, D., LALLOO, F., NAROD, S., RISCH, H., EYFJORD, J., HOPPER, J. et al. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. *Br. J. Cancer* **98** 1457.

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895 https://doi.org/10.1214/aoms/1177728423

BARNETSON, R. A., TENESA, A., FARRINGTON, S. M., NICHOLL, I. D., CETNARSKYJ, R., PORTEOUS, M. E., CAMPBELL, H. and DUNLOP, M. G. (2006). Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *N. Engl. J. Med.* **354** 2751–2763.

BARROW, E., ROBINSON, L., ALDUAIJ, W., SHENTON, A., CLANCY, T., LALLOO, F., HILL, J. and EVANS, D. (2009). Cumulative lifetime incidence of extracolonic cancers in Lynch syndrome: A report of 121 families with proven mutations. *Clin. Genet.* **75** 141–149.

BERNAU, C., RIESTER, M., BOULESTEIX, A.-L., PARMIGIANI, G., HUTTENHOWER, C., WALDRON, L. and TRIPPA, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* **30** i105–i112.

BRAUN, D., YANG, J., GRIFFIN, M., PARMIGIANI, G. and HUGHES, K. S. (2018). A clinical decision support tool to predict cancer risk for commonly tested cancer-related germline mutations. *J. Genet. Couns.* **27** 1187–1199.

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Stat.* **26** 607–616. MR0073894 https://doi.org/10.1214/aoms/1177728420

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794. ACM, New York.

Key words and phrases. Gradient boosting, Risk model, Mendelian model.

CHEN, S. and PARMIGIANI, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.* **25** 1329–1333.

CHEN, S., WANG, W., BROMAN, K. W., KATKI, H. A. and PARMIGIANI, G. (2004). BayesMendel: An R environment for Mendelian risk prediction. *Stat. Appl. Genet. Mol. Biol.* **3** 21. MR2101490 https://doi.org/10.2202/1544-6115.1063

CHEN, S., WANG, W., LEE, S., NAFA, K., LEE, J., ROMANS, K., WATSON, P., GRUBER, S. B., EUHUS, D. et al. (2006). Prediction of germline mutations and cancer risk in the Lynch syndrome. *JAMA* **296** 1479–1487.

COUCH, F. J., DESHANO, M. L., BLACKWOOD, M. A., CALZONE, K., STOPFER, J., CAMPEAU, L., GANGULY, A., REBBECK, T., WEBER, B. L. et al. (1997). BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *N. Engl. J. Med.* **336** 1409–1415.

DEVCAN (2012). DevCan: Probability of Developing or Dying of Cancer Software, Version 6.7.5. Surveillance Research Program, Statistical Methodology and Applications, National Cancer Institute. Available at http://surveillance.cancer.gov/devcan/.

DOWTY, J. G., WIN, A. K., BUCHANAN, D. D., LINDOR, N. M., MACRAE, F. A., CLENDENNING, M., ANTILL, Y. C., THIBODEAU, S. N., CASEY, G. et al. (2013). Cancer risks for MLH1 and MSH2 mutation carriers. *Human Mutat.* **34** 490–497.

ELSTON, R. C. and STEWART, J. (1971). A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21** 523–542.

FAY, M. P. (2004). Estimating age conditional probability of developing disease from surveillance data. *Popul. Health Metr.* **2** 6. https://doi.org/10.1186/1478-7954-2-6

FAY, M. P., PFEIFFER, R., CRONIN, K. A., LE, C. and FEUER, E. J. (2003). Age-conditional probabilities of developing cancer. *Stat. Med.* **22** 1837–1848.

FISHEL, R., LESCOE, M. K., RAO, M., COPELAND, N. G., JENKINS, N. A., GARBER, J., KANE, M. and KOLODNER, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75** 1027–1038.

FLOSSMANN, E., ROTHWELL, P. M. et al. (2007). Effect of aspirin on long-term risk of colorectal cancer: Consistent evidence from randomised and observational studies. *Lancet* **369** 1603–1613.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

FRIEDMAN, J. H. (2002). Stochastic gradient boosting. *Comput. Statist. Data Anal.* **38** 367–378.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* **28** 337–407. MR1790002 https://doi.org/10.1214/aos/1016218223

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HUANG, T., IDOS, G., HONG, C., GRUBER, S., PARMIGIANI, G. and BRAUN, D. (2021). Supplement to "Extending models via gradient boosting: An application to Mendelian models." https://doi.org/10.1214/21-AOAS1482SUPPA, https://doi.org/10.1214/21-AOAS1482SUPPB

IDOS, G., KURIAN, A. W., RICKER, C., STURGEON, D., CULVER, J., KINGHAM, K., KOFF, R., CHUN, N. M., ROWE-TEETER, C. et al. (2018). Promoting breast cancer screening after multiplex genetic panel testing (MGPT) and genetic counseling.

JANSSEN, K. J. M., MOONS, K. G. M., KALKMAN, C. J., GROBBEE, D. E. and VERGOUWE, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *J. Clin. Epidemiol.* **61** 76–86. https://doi.org/10.1016/j.jclinepi.2007.04.018

KASTRINOS, F., IDOS, G. and PARMIGIANI, G. (2018). Prediction models for lynch syndrome. In *Hereditary Colorectal Cancer*: *Genetic Basis and Clinical Implications* (L. Valle, S. B. Gruber and G. Capella, eds.) 281–303. Springer, Cham.

LYNCH, H. T. and SMYRK, T. (1996). Hereditary nonpolyposis colorectal cancer (Lynch syndrome): An updated review. *Cancer*: *Interdiscip. Int. J. Am. Cancer Soc.* **78** 1149–1167.

MARRONI, F., ARETINI, P., D'ANDREA, E., CALIGO, M. A., CORTESI, L., VIEL, A., RICEVUTO, E., MONTAGNA, M., CIPOLLINI, G. et al. (2004). Penetrances of breast and ovarian cancer in a large series of families tested for BRCA1/2 mutations. *Eur. J. Hum. Genet.* **12** 899.

MIYAKI, M., KONISHI, M., TANAKA, K., KIKUCHI-YANOSHITA, R., MURAOKA, M., YASUNO, M., IGARI, T., KOIKE, M., CHIBA, M. et al. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat. Genet.* **17** 271.

MØLLER, P., SEPPÄLÄ, T. T., BERNSTEIN, I., HOLINSKI-FEDER, E., SALA, P., EVANS, D. G., LINDBLOM, A., MACRAE, F., BLANCO, I. et al. (2018). Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: A report from the Prospective Lynch Syndrome Database. *Gut* **67** 1306–1316.

MURPHY, E. and MUTALIK, G. (1969). The application of Bayesian methods in genetic counselling. *Hum. Hered.* **19** 126–151.

NATEKIN, A. and KNOLL, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot.* **7** 21. https://doi.org/10.3389/fnbot.2013.00021

PAPADOPOULOS, N., NICOLAIDES, N. C., WEI, Y.-F., RUBEN, S. M., CARTER, K. C., ROSEN, C. A., HASELTINE, W. A., FLEISCHMANN, R. D., FRASER, C. M. et al. (1994). Mutation of a mutL homolog in hereditary colon cancer. *Science* **263** 1625–1629.

PARMIGIANI, G., CHEN, S., IVERSEN, E. S. JR., FRIEBEL, T. M., FINKELSTEIN, D. M., ANTON-CULVER, H., ZIOGAS, A., WEBER, B. L., EISEN, A. et al. (2007). Validity of models for predicting BRCA1 and BRCA2 mutations. *Ann. Intern. Med.* **147** 441–450.

PLATT, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classifiers* **10** 61–74.

STEYERBERG, E. W., HARRELL, F. E. JR., BORSBOOM, G. J., EIJKEMANS, M., VERGOUWE, Y. and HABBEMA, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54** 774–781.

STEYERBERG, E. W., VICKERS, A. J., COOK, N. R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M. J. and KATTAN, M. W. (2010). Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology* **21** 128.

SU, T.-L., JAKI, T., HICKEY, G. L., BUCHAN, I. and SPERRIN, M. (2018). A review of statistical updating methods for clinical prediction models. *Stat. Methods Med. Res.* **27** 185–197. MR3745663 https://doi.org/10.1177/0962280215626466

VAHTERISTO, P., EEROLA, H., TAMMINEN, A., BLOMQVIST, C. and NEVANLINNA, H. (2001). A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *Br. J. Cancer* **84** 704–708. https://doi.org/10.1054/bjoc.2000.1626

VAN CALSTER, B., NIEBOER, D., VERGOUWE, Y., DE COCK, B., PENCINA, M. J. and STEYERBERG, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* **74** 167–176.

WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259.

ZHANG, Y., BERNAU, C., PARMIGIANI, G. and WALDRON, L. (2020). The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models. *Biostatistics* **21** 253–268. MR4133359 https://doi.org/10.1093/biostatistics/kxy044

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1147–1170
https://doi.org/10.1214/21-AOAS1457

MARKOV-SWITCHING STATE SPACE MODELS FOR UNCOVERING MUSICAL INTERPRETATION

BY DANIEL J. MCDONALD[1], MICHAEL MCBRIDE[2], YUPENG GU[3,*] AND CHRISTOPHER RAPHAEL[3,†]

[1]*Department of Statistics, University of British Columbia, daniel@stat.ubc.ca*
[2]*Department of Statistics, Indiana University, michmcbr@iu.edu*
[3]*School of Informatics, Computing and Engineering, Indiana University, *yupeng.gu@gmail.com; †craphael@indiana.edu*

For concertgoers, musical interpretation is the most important factor in determining whether or not we enjoy a classical performance. Every performance includes mistakes—intonation issues, a lost note, an unpleasant sound—but these are all easily forgotten (or unnoticed) when a performer engages her audience, imbuing a piece with novel emotional content beyond the vague instructions inscribed on the printed page. In this research we use data from the CHARM Mazurka Project—46 professional recordings of Chopin's Mazurka Op. 68 No. 3 by consummate artists—with the goal of elucidating musically interpretable performance decisions. We focus specifically on each performer's use of tempo by examining the interonset intervals of the note attacks in the recording. To explain these tempo decisions, we develop a switching state space model and estimate it by maximum likelihood, combined with prior information gained from music theory and performance practice. We use the estimated parameters to quantitatively describe individual performance decisions and compare recordings. These comparisons suggest methods for informing music instruction, discovering listening preferences and analyzing performances.

REFERENCES

ANDERSON, B. D. O. and MOORE, J. B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115 https://doi.org/10.1111/j.1467-9868.2009.00736.x

ARCOS, J. and MANTARAS, R. L. (2001). An interactive cbr approach for generating expressive music. *J. Appl. Intell.* **21** 115–129.

ARIZA, C. (2005). Navigating the landscape of computer aided algorithmic composition systems: A definition, seven descriptors, and a lexicon of systems and research. In *Proceedings of International Computer Music Conference*.

ARZT, A. and WIDMER, G. (2015). Real-time music tracking using multiple performances as a reference. In *International Society for Music Information Retrieval* (*ISMIR*) 357–363.

BERNSTEIN, L. (2005). *Young People's Concerts*. Amadeus Press, Pompton Plains, NJ.

BISIANI, R. (1992). Beam search. In *Encyclopedia of Artificial Intelligence*, 2nd ed. (S. Shapiro, ed.) Wiley, New York.

BLOCK, B. A., JONSEN, I. D., JORGENSEN, S. J., WINSHIP, A. J., SHAFFER, S. A., BOGRAD, S. J., HAZEN, E. L., FOLEY, D. G., BREED, G. et al. (2011). Tracking apex marine predator movements in a dynamic ocean. *Nature* **475** 86.

BOULANGER-LEWANDOWSKI, N., BENGIO, Y. and VINCENT, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *Proceedings of the* 29*th International Conference on Machine Learning*.

BRESIN, R., FRIBERG, A. and SUNDBERG, J. (2002). Director musices: The KTH performance rules system. In *Proceedings of SIGMUS*-46.

BURKHOLDER, J. P., GROUT, D. J. and PALISCA, C. V. (2014). *A History of Western Music*, 9th ed. Norton, New York.

CHARM (2009). Centre for the History and Analysis of Recorded Music. Online; accessed 12 March 2019.

Key words and phrases. Classification and clustering, Kalman filter, hidden Markov model.

COLLINS, N. (2016). A funny thing happened on the way to the formula: Algorithmic composition for musical theater. *Comput. Music J.* **40** 41–57.
CONT, A. (2010). A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 974–987.
CONT, A., SCHWARZ, D., SCHNELL, N. and RAPHAEL, C. (2007). Evaluation of real-time audio-to-score alignment. In *International Symposium on Music Information Retrieval* (*ISMIR*).
COOK, N. (2013). *Beyond the Score*: *Music as Performance*. Oxford Univ. Press, Oxford.
CRAVEN, P. and WAHBA, G. (1978). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.
DANNENBERG, R. (1985). An on-line algorithm for real-time accompaniment. In *Proceedings of the* 1984 *International Computer Music Conference* 193–198. International Computer Music Association.
DANNENBERG, R. B. and RAPHAEL, C. (2006). Music score alignment and computer accompaniment. *Commun. ACM* **49** 38–43.
DROR, G., KOENIGSTEIN, N., KOREN, Y. and WEIMER, M. (2012). The yahoo! Music dataset and KDD-Cup'11. In *KDD Cup* 8–18.
DUDOIT, S. and FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3** research0036.1.
DURBIN, J. and KOOPMAN, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84** 669–684. MR1603944 https://doi.org/10.1093/biomet/84.3.669
DURBIN, J. and KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. *Oxford Statistical Science Series* **24**. Oxford Univ. Press, Oxford. MR1856951
EARIS, A. (2007). An algorithm to extract expressive timing and dynamics from piano recordings. *Music. Sci.* **11** 155–182.
EARIS, A. (2009). Mazurka in F Major, Op. 68, No. 3. accessed 12 March 2019.
EDDELBUETTEL, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York.
FEARNHEAD, P. and CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 887–899. MR2017876 https://doi.org/10.1111/1467-9868.00421
FLOSSMAN, S., GRACHTEN, M. and WIDMER, G. (2012). Expressive performance rendering with probabilistic models. In *Guide to Computing for Expressive Music Performance* (A. Kirke and E. Miranda, eds.) Springer, Berlin.
FLOSSMANN, S., GRACHTEN, M. and WIDMER, G. (2013). Expressive performance rendering with probabilistic models. In *Guide to Computing for Expressive Music Performance* 75–98. Springer, Berlin.
FORSÉN, S., GRAY, H. B., LINDGREN, L. K. O. and GRAY, S. B. (2013). Was something wrong with Beethoven's metronome? *Notices Amer. Math. Soc.* **60** 1146–1153. MR3113276 https://doi.org/10.1090/noti1044
FOX, E. B., SUDDERTH, E. B., JORDAN, M. I. and WILLSKY, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Stat.* **5** 1020–1056. MR2840185 https://doi.org/10.1214/10-AOAS395
FUH, C.-D. (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34** 2026–2068. MR2283726 https://doi.org/10.1214/009053606000000614
GHAHRAMANI, Z. and HINTON, G. E. (2000). Variational learning for switching state-space models. *Neural Comput.* **12** 831–864.
GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. MR0533250 https://doi.org/10.2307/1268518
GRINDLAY, G. and HELMBOLD, D. (2006). Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Mach. Learn.* **65** 361–387.
GU, Y. and RAPHAEL, C. (2012). Modeling piano interpretation using switching Kalman filter. In *International Society for Music Information Retrieval* (*ISMIR*) 145–150.
HADJERES, G., PACHET, F. and NIELSEN, F. (2017). DeepBach: A steerable model for bach chorales generation. In *Proceedings of the* 34*th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 1362–1371. PMLR, Sydney, Australia.
HAMILTON, J. D. (2011). Calling recessions in real time. *Int. J. Forecast.* **27** 1006–126.
HARVEY, A. C. (1990). *Forecasting*, *Structural Time Series Models and the Kalman Filter*. Cambridge Univ. Press, Cambridge.
KALLBERG, J. (1996). *Chopin at the Boundaries*: *Sex*, *History*, *and Musical Genre*. Harvard Univ. Press, Harvard.
KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45. MR3931993
KIM, C.-J. (1994). Dynamic linear models with Markov-switching. *J. Econometrics* **60** 1–22. MR1247815 https://doi.org/10.1016/0304-4076(94)90036-1
KIM, C. J. and NELSON, C. R. (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *Rev. Econ. Stat.* **80** 188–201.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.* **82** 1032–1063. MR0922169

KITAGAWA, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.* **5** 1–25. MR1380850 https://doi.org/10.2307/1390750

KOYAMA, S., CASTELLANOS PÉREZ-BOLDE, L., SHALIZI, C. R. and KASS, R. E. (2010). Approximate methods for state-space models. *J. Amer. Statist. Assoc.* **105** 170–180. MR2656047 https://doi.org/10.1198/jasa.2009.tm08326

LANG, M., BISCHL, B. and SURMANN, D. (2017). batchtools: Tools for R to work on batch systems. *J. Open Sour. Softw.* **2** 135.

LANG, D. and FREITAS, N. D. (2005). Beat tracking the graphical model way. In *Advances in Neural Information Processing Systems* 745–752. MIT press, Cambridge, MA.

MAEZAWA, A. (2019). Deep linear autoregressive model of interpretable prediction of expressive tempo. In *Proceedings of the* 16*th Sound and Music Computing Conference*.

MCDONALD, D. J, MCBRIDE, M., GU, Y. and RAPHAEL, C. (2021). Supplement to "Markov-Switching State Space Models for Uncovering Musical Interpretation." https://doi.org/10.1214/21-AOAS1457SUPPA, https://doi.org/10.1214/21-AOAS1457SUPPB, https://doi.org/10.1214/21-AOAS1457SUPPC

MCFEE, B. and LANCKRIET, G. (2011). Learning multi-modal similarity. *J. Mach. Learn. Res.* **12** 491–523. MR2783175

MEAD, A. (2007). On tempo relations. *Perspect. New Music* **45** 64–108.

PATTERSON, T. A., THOMAS, L., WILCOX, C., OVASKAINEN, O. and MATTHIOPOULOS, J. (2008). State–space models of individual animal movement. *Trends Ecol. Evol.* **23** 87–94.

R CORE TEAM (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RAPHAEL, C. (2002). A hybrid graphical model for rhythmic parsing. *Artificial Intelligence* **137** 217–238.

RAPHAEL, C. (2010). Music plus one and machine learning. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*-10) (J. Fürnkranz and T. Joachims, eds.) 21–28.

RAUCH, H. E., TUNG, F. and STRIEBEL, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA J.* **3** 1445–1450. MR0181489 https://doi.org/10.2514/3.3166

REN, L., DUNSON, D., LINDROTH, S. and CARIN, L. (2010). Dynamic nonparametric Bayesian models for analysis of music. *J. Amer. Statist. Assoc.* **105** 458–472. MR2724839 https://doi.org/10.1198/jasa.2009.ap08497

ROBERTS, A., HAWTHORNE, C. and SIMON, I. (2018). Magenta.js: A JavaScript API for augmenting creativity with deep learning. In *Joint Workshop on Machine Learning for Music* (*ICML*).

ROBERTS, A., ENGEL, J., RAFFEL, C., HAWTHORNE, C. and ECK, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 4364–4373. PMLR, Stockholmsmässan, Stockholm, Sweden.

SCHEDL, M., GÓMEZ, E., URBANO, J. et al. (2014). Music information retrieval: Recent developments and applications. *Found. Trends Inf. Retr.* **8** 127–261.

STOWELL, D. and CHEW, E. (2012). Bayesian MAP estimation of piecewise arcs in tempo time series. In *Proceedings of Computer Music Multidisciplinary Research*.

STURM, B. L., BEN-TAL, O., MONAGHAN, Ú., COLLINS, N., HERREMANS, D., CHEW, E., HADJERES, G., DERUTY, E. and PACHET, F. (2019). Machine learning research that matters for music creation: A case study. *J. New Music Res.* **48** 36–55.

THICKSTUN, J., HARCHAOUI, Z. and KAKADE, S. M. (2017). Learning features of music from scratch. In *International Conference on Learning Representations* (*ICLR*).

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

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 https://doi.org/10.1111/1467-9868.00293

VAN DEN OORD, A., DIELEMAN, S. and SCHRAUWEN, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2643–2651. Curran Associates, Red Hook.

VERCOE, B. (1984). The synthetic performer in the context of live performance. In *Proceedings of the* 1984 *International Computer Music Conference* 199–200. International Computer Music Association.

WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128

WHITELEY, N., ANDRIEU, C. and DOUCET, A. (2010). Efficient Bayesian Inference for Switching State-Space Models using Discrete Particle Markov Chain Monte Carlo Methods Technical Report No. 10:04 Bristol Univ.
WICKHAM, H. (2016). *Ggplot2*: *Elegant Graphics for Data Analysis*, 2nd ed. Springer, Berlin.
WICKHAM, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.
WIDMER, G., FLOSSMANN, S. and GRACHTEN, M. (2009). YQX plays chopin. *AI Mag*. **30** 35.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1171–1193
https://doi.org/10.1214/21-AOAS1455

IDENTIFYING THE RECURRENCE OF SLEEP APNEA USING A HARMONIC HIDDEN MARKOV MODEL

BY BENIAMINO HADJ-AMAR[1,*], BÄRBEL FINKENSTÄDT[1,†], MARK FIECAS[2] AND ROBERT HUCKSTEPP[3]

[1]*Department of Statistics, University of Warwick,* [*]*Beniamino.Hadj-Amar@rice.edu;* [†]*B.F.Finkenstadt@warwick.ac.uk*
[2]*School of Public Health, Division of Biostatistics, University of Minnesota, mfiecas@umn.edu*
[3]*School of Life Sciences, University of Warwick, R.Huckstepp@warwick.ac.uk*

We propose to model time-varying periodic and oscillatory processes by means of a hidden Markov model where the states are defined through the spectral properties of a periodic regime. The number of states is unknown along with the relevant periodicities, the role and number of which may vary across states. We address this inference problem by a Bayesian nonparametric hidden Markov model, assuming a sticky hierarchical Dirichlet process for the switching dynamics between different states while the periodicities characterizing each state are explored by means of a transdimensional Markov chain Monte Carlo sampling step. We develop the full Bayesian inference algorithm and illustrate the use of our proposed methodology for different simulation studies as well as an application related to respiratory research which focuses on the detection of apnea instances in human breathing traces.

REFERENCES

ADAK, S. (1998). Time-dependent spectral analysis of nonstationary time series. *J. Amer. Statist. Assoc.* **93** 1488–1501. MR1666643 https://doi.org/10.2307/2670062

ALBERT, J. H. and CHIB, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Bus. Econom. Statist.* **11** 1–15.

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 https://doi.org/10.1007/BFb0099421

ANCOLI-ISRAEL, S., KLAUBER, M. R., BUTTERS, N., PARKER, L. and KRIPKE, D. F. (1991). Dementia in institutionalized elderly: Relation to sleep apnea. *J. Amer. Geriatr. Soc.* **39** 258–263.

ANDRIEU, C. and DOUCET, A. (1999). Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. Signal Process.* **47** 2667–2676.

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115 https://doi.org/10.1111/j.1467-9868.2009.00736.x

BAUM, L. E. and EAGON, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73** 360–363. MR0210217 https://doi.org/10.1090/S0002-9904-1967-11751-8

BAUM, L. E. and PETRIE, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37** 1554–1563. MR0202264 https://doi.org/10.1214/aoms/1177699147

BEAL, M. J., GHAHRAMANI, Z. and RASMUSSEN, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems* 577–584.

BERNARDO, J.-M. and SMITH, A. F. M. (2009). *Bayesian Theory*. Wiley, Chichester. MR1274699 https://doi.org/10.1002/9780470316870

BERRY, R. B., BROOKS, R., GAMALDO, C., HARDING, S. M., LLOYD, R. M., QUAN, S. F., TROESTER, M. T. and VAUGHN, B. V. (2017). AASM scoring manual updates for 2017 (version 2.4). *J. Clin. Sleep. Med.* **13** 665–666. https://doi.org/10.5664/jcsm.6576

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BRUCE, S. A., HALL, M. H., BUYSSE, D. J. and KRAFTY, R. T. (2018). Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series. *Biometrics* **74** 260–269. MR3777946 https://doi.org/10.1111/biom.12719

Key words and phrases. Sleep apnea, time-varying frequencies, reversible-jump MCMC, Bayesian nonparametrics, hierarchical Dirichlet process.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. *Springer Series in Statistics*. Springer, New York. MR2159833
CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450 https://doi.org/10.2307/2669477
COHEN, M. X. (2014). *Analyzing Neural Time Series Data*: *Theory and Practice*. MIT Press, Cambridge.
COOKE, J. R., AYALON, L., PALMER, B. W., LOREDO, J. S., COREY-BLOOM, J., NATARAJAN, L., LIU, L. and ANCOLI-ISRAEL, S. (2009). Sustained use of CPAP slows deterioration of cognition, sleep, and mood in patients with Alzheimer's disease and obstructive sleep apnea: A preliminary study. *J. Clin. Sleep. Med.* **5** 305–309.
DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916 https://doi.org/10.1214/aos/1034276620
DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101** 223–239. MR2268041 https://doi.org/10.1198/016214505000000745
DEWAN, N. A., NIETO, F. J. and SOMERS, V. K. (2015). Intermittent hypoxemia and OSA: Implications for comorbidities. *Chest* **147** 266–274. https://doi.org/10.1378/chest.14-0500
EPHRAIM, Y. and MERHAV, N. (2002). Hidden Markov processes. *IEEE Trans. Inf. Theory* **48** 1518–1569. MR1909472 https://doi.org/10.1109/TIT.2002.1003838
FOX, E. B., SUDDERTH, E. B., JORDAN, M. I. and WILLSKY, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *Ann. Appl. Stat.* **5** 1020–1056. MR2840185 https://doi.org/10.1214/10-AOAS395
FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*. Springer, New York. MR2265601
GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677
GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711
GUÉDON, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *J. Comput. Graph. Statist.* **12** 604–639. MR2002638 https://doi.org/10.1198/1061860032030
HADJ-AMAR, B., RAND, B. F., FIECAS, M., LÉVI, F. and HUCKSTEPP, R. (2020). Bayesian model search for nonstationary periodic time series. *J. Amer. Statist. Assoc.* **115** 1320–1335. MR4143468 https://doi.org/10.1080/01621459.2019.1623043
HADJ-AMAR, B., FINKENSTÄDT, B., FIECAS, M. and HUCKSTEPP, R. (2021). Supplement to "Identifying the recurrence of sleep apnea using a harmonic hidden Markov model." https://doi.org/10.1214/21-AOAS1455SUPP
HEINZER, R., VAT, S., MARQUES-VIDAL, P., MARTI-SOLER, H., ANDRIES, D., TOBBACK, N., MOOSER, V., PREISIG, M., MALHOTRA, A. et al. (2015). Prevalence of sleep-disordered breathing in the general population: The HypnoLaus study. *The Lancet Respiratory Medicine* **3** 310–318.
HUANG, Q., COHEN, D., KOMARZYNSKI, S., LI, X.-M., INNOMINATO, P., LÉVI, F. and FINKENSTÄDT, B. (2018). Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *J. R. Soc. Interface* **15** 20170885.
HURN, M., JUSTEL, A. and ROBERT, C. P. (2003). Estimating mixtures of regressions. *J. Comput. Graph. Statist.* **12** 55–79. MR1977206 https://doi.org/10.1198/1061860031329
IGNATOV, T. (1982). A constant arising in the asymptotic theory of symmetric groups, and Poisson–Dirichlet measures. *Theory Probab. Appl.* **27** 136–147.
ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. MR1926065 https://doi.org/10.2307/3315951
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 https://doi.org/10.1214/088342305000000016
JASRA, A., STEPHENS, D. A. and HOLMES, C. C. (2007). On population-based simulation for static inference. *Stat. Comput.* **17** 263–279. MR2405807 https://doi.org/10.1007/s11222-007-9028-9
JASRA, A., DOUCET, A., STEPHENS, D. A. and HOLMES, C. C. (2008). Interacting sequential Monte Carlo samplers for trans-dimensional simulation. *Comput. Statist. Data Anal.* **52** 1765–1791. MR2418470 https://doi.org/10.1016/j.csda.2007.09.009
JOHNSON, M. J. and WILLSKY, A. S. (2013). Bayesian nonparametric hidden semi-Markov models. *J. Mach. Learn. Res.* **14** 673–701. MR3033344
JUANG, B.-H. and RABINER, L. (1985). Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. Acoust. Speech Signal Process.* **33** 1404–1413.
KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

KIVINEN, J. J., SUDDERTH, E. B. and JORDAN, M. I. (2007). Learning multiscale representations of natural scenes using Dirichlet processes. In 2007 *IEEE* 11*th International Conference on Computer Vision* 1–8. IEEE, New York.

KOMARZYNSKI, S., HUANG, Q., INNOMINATO, P. F., MAURICE, M., ARBAUD, A., BEAU, J., BOUCHAHDA, M., ULUSAKARYA, A., BEAUMATIN, N. et al. (2018). Relevance of a mobile Internet platform for capturing inter- and intrasubject variabilities in circadian coordination during daily routine: Pilot study. *J. Med. Internet Res.* **20** e204. https://doi.org/10.2196/jmir.9779

KRAUCHI, K. and WIRZ-JUSTICE, A. (1994). Circadian rhythm of heat production, heart rate, and skin and core temperature under unmasking conditions in men. *Am. J. Physiol., Regul. Integr. Comp. Physiol.* **267** R819–R829.

KROGH, A., BROWN, M., MIAN, I. S., SJÖLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* **235** 1501–1531.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22** 79–86. MR0039968 https://doi.org/10.1214/aoms/1177729694

LANGROCK, R., SWIHART, B. J., CAFFO, B. S., PUNJABI, N. M. and CRAINICEANU, C. M. (2013). Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Stat. Med.* **32** 3342–3356. MR3074361 https://doi.org/10.1002/sim.5747

LIANG, F. and WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *J. Amer. Statist. Assoc.* **96** 653–666. MR1946432 https://doi.org/10.1198/016214501753168325

MALIK, M. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Annals of Noninvasive Electrocardiology* **1** 151–181.

MARIN, J.-M., MENGERSEN, K. and ROBERT, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking*: *Modeling and Computation*. *Handbook of Statist.* **25** 459–507. Elsevier/North-Holland, Amsterdam. MR2490536 https://doi.org/10.1016/S0169-7161(05)25016-2

MENG, X.-L. and SCHILLING, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11** 552–586. MR1938446 https://doi.org/10.1198/106186002457

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. MR1994729 https://doi.org/10.1214/aos/1056562461

OMBAO, H. C., RAZ, J. A., VON SACHS, R. and MALOW, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc.* **96** 543–560. MR1946424 https://doi.org/10.1198/016214501753168244

PAPASTAMOULIS, P. (2016). Label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, *Code Snippets* **69** 1–24.

PAPASTAMOULIS, P. and ILIOPOULOS, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J. Comput. Graph. Statist.* **19** 313–331. MR2758306 https://doi.org/10.1198/jcgs.2010.09008

PAZ, J. C. and WEST, M. P. (2013). *Acute Care Handbook for Physical Therapists*. Elsevier Health Sciences, Elsevier.

PEKER, Y., HEDNER, J., NORUM, J., KRAICZI, H. and CARLSON, J. (2002). Increased incidence of cardiovascular disease in middle-aged men with obstructive sleep apnea: A 7-year follow-up. *Am. J. Respir. Crit. Care Med.* **166** 159–165.

PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448 https://doi.org/10.1007/BF01205234

PITMAN, J. (1996). Blackwell–Macqueen urn scheme. *Statistics*, *Probability, and Game Theory*: *Papers in Honor of David Blackwell* **30** 245.

PITMAN, J. (2002). Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition. *Combin. Probab. Comput.* **11** 501–514. MR1930355 https://doi.org/10.1017/S0963548302005163

PRIESTLEY, M. B. (1965). Evolutionary spectra and non-stationary processes.(With discussion). *J. Roy. Statist. Soc. Ser. B* **27** 204–237. MR0199886

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

RASMUSSEN, C. E. and GHAHRAMANI, Z. (2002). Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems* 881–888.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930 https://doi.org/10.1137/1026034

RODRÍGUEZ, C. E. and WALKER, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *J. Comput. Graph. Statist.* **23** 25–45. MR3173759 https://doi.org/10.1080/10618600.2012.735624

ROSEN, O., STOFFER, D. S. and WOOD, S. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *J. Amer. Statist. Assoc.* **104** 249–262. MR2504376 https://doi.org/10.1198/jasa.2009.0118

ROSEN, O., WOOD, S. and STOFFER, D. S. (2012). AdaptSPEC: Adaptive spectral estimation for nonstationary time series. *J. Amer. Statist. Assoc.* **107** 1575–1589. MR3036417 https://doi.org/10.1080/01621459.2012.716340

RUEHLAND, W. R., ROCHFORD, P. D., O'DONOGHUE, F. J., PIERCE, R. J., SINGH, P. and THORNTON, A. T. (2009). The new AASM criteria for scoring hypopneas: Impact on the apnea hypopnea index. *Sleep* **32** 150–157.

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

SHUMWAY, R. H. and STOFFER, D. S. (2017). *Time Series Analysis and Its Applications*: *With R Examples*, 4th ed. *Springer Texts in Statistics*. Springer, Cham. MR3642322 https://doi.org/10.1007/978-3-319-52452-8

STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293 https://doi.org/10.1111/1467-9868.00265

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

TERAN-SANTOS, J., JIMENEZ-GOMEZ, A., CORDERO-GUEVARA, J. and BURGOS-SANTANDER, C. G. (1999). The association between sleep apnea and the risk of traffic accidents. *N. Engl. J. Med.* **340** 847–851.

TRIPURANENI, N., GU, S. S., GE, H. and GHAHRAMANI, Z. (2015). Particle Gibbs for infinite hidden Markov models. In *Advances in Neural Information Processing Systems* 2395–2403.

VAN GAEL, J., SAATCI, Y., TEH, Y. W. and GHAHRAMANI, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the* 25*th International Conference on Machine Learning* 1088–1095. ACM, New York.

WALKER, S. G. (2010). Bayesian nonparametric methods: Motivation and ideas. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math.* **28** 22–34. Cambridge Univ. Press, Cambridge. MR2722988 https://doi.org/10.1017/CBO9780511802478.002

WEST, M., PRADO, R. and KRYSTAL, A. D. (1999). Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *J. Amer. Statist. Assoc.* **94** 375–387.

WHITTLE, P. (1957). Curve and periodogram smoothing. *J. Roy. Statist. Soc. Ser. B* **19** 38–47 (discussion 47–63). MR0092331

YAGGI, H. K., CONCATO, J., KERNAN, W. N., LICHTMAN, J. H., BRASS, L. M. and MOHSENIN, V. (2005). Obstructive sleep apnea as a risk factor for stroke and death. *N. Engl. J. Med.* **353** 2034–2041.

YAGHOUBY, F. and SUNDERAM, S. (2015). Quasi-supervised scoring of human sleep in polysomnograms using augmented input variables. *Comput. Biol. Med.* **59** 54–63. https://doi.org/10.1016/j.compbiomed.2015.01.012

YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G. O. and HOLMES, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 37–57. MR2797735 https://doi.org/10.1111/j.1467-9868.2010.00756.x

YOUNG, T., PEPPARD, P. E. and GOTTLIEB, D. J. (2002). Epidemiology of obstructive sleep apnea: A population health perspective. *Am. J. Respir. Crit. Care Med.* **165** 1217–1239.

YU, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence* **174** 215–243. MR2724430 https://doi.org/10.1016/j.artint.2009.11.011

ZHOU, Y., JOHANSEN, A. M. and ASTON, J. A. D. (2016). Toward automatic model comparison: An adaptive sequential Monte Carlo approach. *J. Comput. Graph. Statist.* **25** 701–726. MR3533634 https://doi.org/10.1080/10618600.2015.1060885

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1194–1219
https://doi.org/10.1214/20-AOAS1438

TARGETED SMOOTH BAYESIAN CAUSAL FORESTS: AN ANALYSIS OF HETEROGENEOUS TREATMENT EFFECTS FOR SIMULTANEOUS VS. INTERVAL MEDICAL ABORTION REGIMENS OVER GESTATION

BY JENNIFER E. STARLING[1], JARED S. MURRAY[2,*], PATRICIA A. LOHR[3], ABIGAIL R. A. AIKEN[4], CARLOS M. CARVALHO[2,†] AND JAMES G. SCOTT[2,‡]

[1]*Department of Statistics and Data Sciences, University of Texas at Austin, jstarling@utexas.edu*

[2]*Department of Statistics and Data Sciences and McCombs School of Business, University of Texas at Austin,* [*]*jared.murray@mccombs.utexas.edu;* [†]*carlos.carvalho@mccombs.utexas.edu;* [‡]*james.scott@mccombs.utexas.edu*

[3]*British Pregnancy Advisory Service, patricia.lohr@bpas.org*

[4]*Lyndon B. Johnson School of Public Affairs, University of Texas at Austin, araa2@utexas.edu*

We introduce Targeted Smooth Bayesian Causal Forests (tsBCF), a nonparametric Bayesian approach for estimating heterogeneous treatment effects which vary smoothly over a single covariate in the observational data setting. The tsBCF method induces smoothness by parameterizing terminal tree nodes with smooth functions and allows for separate regularization of treatment effects vs. prognostic effect of control covariates. Smoothing parameters for prognostic and treatment effects can be chosen to reflect prior knowledge or tuned in a data-dependent way.

We use tsBCF to analyze a new clinical protocol for early medical abortion. Our aim is to assess the relative effectiveness of simultaneous vs. interval administration of mifepristone and misoprostol over the first nine weeks of gestation. Our analysis yields important clinical insights into how to best counsel patients seeking early medical abortion, where understanding even small differences in relative effectiveness can yield dramatic returns to public health. The model reflects our expectation that the treatment effect varies smoothly over gestation but not necessarily over other covariates. We demonstrate the performance of the tsBCF method on benchmarking experiments. Software for tsBCF is available at https://github.com/jestarling/tsbcf/ and in the Supplementary Material (Starling (2020)).

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ATHEY, S. and WAGER, S. (2019). Estimating treatment effects with Causal Forests: An application. Available at arXiv:1902.07409.

CAMERON, S., GLASIER, A., JOHNSTON, A., DEWART, H. and CAMPBELL, A. (2015). Can women determine the success of early medical termination of pregnancy themselves? *Contraception* **91** 6–11.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CREININ, M. and CHEN, M. (2016). Medical abortion reporting of efficacy: The MARE guidelines. *Contraception* **94** 97–103.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GOEL, A., MITTAL, S., TANEJA, B. K., SINGAL, N. and ATTRI, S. (2011). Simultaneous administration of mifepristone and misoprostol for early termination of pregnancy: A randomized controlled trial **283** 1409. https://doi.org/10.1007/s00404-011-1881-2

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 https://doi.org/10.1198/016214508000000689

Key words and phrases. Bayesian additive regression tree, causal inference, regularization, Gaussian process, heterogeneous treatment effects.

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

HAHN, P. R., CARVALHO, C. M., PUELZ, D. and HE, J. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Anal.* **13** 163–182. MR3737947 https://doi.org/10.1214/16-BA1044

HEFFER, G. (2020). Coronavirus: Government relaxes abortion rules during COVID-19 crisis. *Sky News*.

HERNÁNDEZ, B., RAFTERY, A. E., PENNINGTON, S. R. and PARNELL, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Stat. Comput.* **28** 869–890. MR3766048 https://doi.org/10.1007/s11222-017-9767-1

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

IMBENS, G. and RUBIN, D. (2015). Causal inference in statistics, social, and biomedical sciences. *Cambridge University Press*.

LI, Y.-T., CHEN, F.-M., CHEN, T.-H., LI, S.-C., CHEN, M.-L. and KUO, T.-C. (2006). Concurrent use of mifepristone and misoprostol for early medical abortion **45** 325–328. https://doi.org/10.1016/S1028-4559(09)60252-7

LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. MR3832214 https://doi.org/10.1080/01621459.2016.1264957

LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1087–1110. MR3874311 https://doi.org/10.1111/rssb.12293

LOGAN, B. R., SPARAPANI, R., MCCULLOCH, R. E. and LAUD, P. W. (2017). Subgroup finding via Bayesian additive regression trees. *Stat. Methods Med. Res.* **0** 1–15.

LOHR, P. A., STARLING, J. E., SCOTT, J. G. and AIKEN, A. R. A. (2018). Simultaneous compared with inverval medical abortion regimens where home use is restricted. *Obstet. Gynecol.*.

MCFADDEN, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (P. Zarembka, ed.) 105–142. Academic Press, San Diego.

MURRAY, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. Preprint. Available at arXiv:1701.01503.

ROYAL COLLEGE OF OBSTETRICIANS AND GYNAECOLOGISTS (2011). *The Care of Women Requesting Induced Abortion*. RCOG, London, UK.

PRATOLA, M. T., CHIPMAN, H. A., GATTIKER, J. R., HIGDON, D. M., MCCULLOCH, R. and RUST, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist.* **23** 830–852. MR3224658 https://doi.org/10.1080/10618600.2013.841584

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 https://doi.org/10.1093/biomet/70.1.41

SIVAGANESAN, S., MÜLLER, P. and HUANG, B. (2017). Subgroup finding via Bayesian additive regression trees. *Stat. Med.* **36** 2391–2403. MR3660139 https://doi.org/10.1002/sim.7276

SPARAPANI, R. A., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Stat. Med.* **35** 2741–2753. MR3513715 https://doi.org/10.1002/sim.6893

STARLING, J. (2020). tsbcf: Targeted Smooth Bayesian Causal Forests. R package version 1.1.0.

STARLING, J. E., MURRAY, J. S., CARVALHO, C. M., BUKOWSKI, R. K. and SCOTT, J. G. (2020). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *Ann. Appl. Stat.* **14** 28–50. MR4085082 https://doi.org/10.1214/19-AOAS1268

STARLING, J. E, MURRAY, J. S, LOHR, P. A, AIKEN, A. R, CARVALHO, C. M and SCOTT, J. G (2021). Supplement to "Targeted Smooth Bayesian Causal Forests: An analysis of heterogeneous treatment effects for simultaneous vs. interval medical abortion regimens over gestation." https://doi.org/10.1214/20-AOAS1438SUPP

TIBSHIRANI, R., ATHEY, S., FRIEDBERG, R., HADAD, V., MINER, L., WAGER, S. and WRIGHT, M. (2018). grf: Generalized Random Forests. Available at https://github.com/grf-labs/grf. R package version 1.2.0.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

WOODY, S., CARVALHO, C. and MURRAY, J. (2019). Model interpretation through lower-dimensional posterior summarization. Preprint. Available at arXiv:1905.07103v3 [stat.ME].

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1220–1246
https://doi.org/10.1214/21-AOAS1487

STABILIZING VARIABLE SELECTION AND REGRESSION

BY NIKLAS PFISTER[1,*], EVAN G. WILLIAMS[2], JONAS PETERS[1,†], RUEDI AEBERSOLD[3] AND PETER BÜHLMANN[4]

[1]*Department of Mathematical Sciences, University of Copenhagen,* [*]*np@math.ku.dk;* [†]*jonas.peters@math.ku.dk*
[2]*Luxembourg Centre for Systems Biomedicine, University of Luxembourg, evan.williams@uni.lu*
[3]*Institute of Molecular Systems Biology, ETH Zürich, aebersold@imsb.biol.ethz.ch*
[4]*Seminar for Statistics, ETH Zürich, buehlmann@stat.math.ethz.ch*

We consider regression in which one predicts a response Y with a set of predictors X across different experiments or environments. This is a common setup in many data-driven scientific fields, and we argue that statistical inference can benefit from an analysis that takes into account the distributional changes across environments. In particular, it is useful to distinguish between stable and unstable predictors, that is, predictors which have a fixed or a changing functional dependence on the response, respectively. We introduce stabilized regression which explicitly enforces stability and thus improves generalization performance to previously unseen environments. Our work is motivated by an application in systems biology. Using multiomic data, we demonstrate how hypothesis generation about gene function can benefit from stabilized regression. We believe that a similar line of arguments for exploiting heterogeneity in data can be powerful for many other applications as well. We draw a theoretical connection between multi-environment regression and causal models which allows to graphically characterize stable vs. unstable functional dependence on the response. Formally, we introduce the notion of a stable blanket which is a subset of the predictors that lies between the direct causal predictors and the Markov blanket. We prove that this set is optimal in the sense that a regression based on these predictors minimizes the mean squared prediction error, given that the resulting regression generalizes to unseen new environments.

REFERENCES

ALDRICH, J. (1989). Autonomy. *Oxf. Econ. Pap.* **41** 15–34.

BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BÜHLMANN, P. (2020). Invariance, causality and robustness: 2018 Neyman Lecture. *Statist. Sci.* **35** 404–426. MR4148216 https://doi.org/10.1214/19-STS721

BURNHAM, K. and ANDERSON, D. (1998). Practical use of the information-theoretic approach. In *Model Selection and Inference* 75–117. Springer.

CANNINGS, T. I. and SAMWORTH, R. J. (2017). Random-projection ensemble classification. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 959–1035. With discussions and a reply by the authors. MR3689307 https://doi.org/10.1111/rssb.12228

CHOW, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605. MR0141193 https://doi.org/10.2307/1910133

CONSTANTINOU, P. and DAWID, A. P. (2017). Extended conditional independence and applications in causal inference. *Ann. Statist.* **45** 2618–2653. MR3737904 https://doi.org/10.1214/16-AOS1537

ČUKLINA, J., LEE, C. H., WILLIAMS, E. G., SAJIC, T., COLLINS, B. C., RODRIGUEZ MARTINEZ, M., SHARMA, V. S., WENDT, F., GOETZE, S., KEELE, G. R. et al. (2021). Molecular systems biology. Batch effects in large-scale proteomics studies: diagnostics and correction.

DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* **70** 161–189.

DUTKOWSKI, J., KRAMER, M., SURMA, M. A., BALAKRISHNAN, R., CHERRY, J. M., KROGAN, N. J. and IDEKER, T. (2013). A gene ontology inferred from molecular networks. *Nat. Biotechnol.* **31** 38–45.

Key words and phrases. Regression, variable selection, causality, multiomic data.

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F. et al. (2017). The reactome pathway knowledgebase. *Nucleic Acids Res*. **46(D1)** D649–D655, 11. https://doi.org/10.1093/nar/gkx1132

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FRANCESCONI, M., REMONDINI, D., NERETTI, N., SEDIVY, J. M., COOPER, L. N., VERONDINI, E., MILANESI, L. and CASTELLANI, G. (2008). Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinform*. **9** 9.

GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M. and LEMPITSKY, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res*. **17** Paper No. 59, 35. MR3504619

HAAVELMO, T. (1944). The probability approach in econometrics. *Econometrica* **12 S** 118. MR0010953 https://doi.org/10.2307/1906935

HEINZE-DEML, C. and MEINSHAUSEN, N. (2021). Conditional variance penalties and domain shift robustness. *Mach. Learn*. **110** 303–348. MR4207502 https://doi.org/10.1007/s10994-020-05924-1

HEINZE-DEML, C., PETERS, J. and MEINSHAUSEN, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference* **6**.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: A tutorial. *Statist. Sci*. **14** 382–417. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. MR1765176 https://doi.org/10.1214/ss/1009212519

HOOVER, K. D. (1990). The logic of causal inference. *Econ. Philos*. **6** 207–234.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28** 27–30. https://doi.org/10.1093/nar/28.1.27

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **72** 417–473. MR2758523 https://doi.org/10.1111/j.1467-9868.2010.00740.x

PAN, S., TSANG, I., KWOK, J. and YANG, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw*. **22** 199–210.

PEARL, J. (2009). *Causality*: *Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PELLET, J.-P. and ELISSEEFF, A. (2008). Using Markov blankets for causal structure learning. *J. Mach. Learn. Res*. **9** 1295–1342. MR2426044

PERRONE, M. and COOPER, L. (1992). When networks disagree: Ensemble methods for hybrid neural networks. Technical report, Brown Univ., Providence RI, Institute for Brain and Neural Systems.

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **78** 947–1012. With comments and a rejoinder. MR3557186 https://doi.org/10.1111/rssb.12167

PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference*: *Foundations and Learning Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

PFISTER, N., BAUER, S. and PETERS, J. (2019). Learning stable and predictive structures in kinetic systems. *Proc. Natl. Acad. Sci. USA* **116** 25405–25411. MR4047351 https://doi.org/10.1073/pnas.1905688116

PFISTER, N., BÜHLMANN, P. and PETERS, J. (2019). Invariant causal prediction for sequential data. *J. Amer. Statist. Assoc*. **114** 1264–1276. MR4011778 https://doi.org/10.1080/01621459.2018.1491403

PFISTER, N., WILLIAMS, E. G., AEBERSOLD, R. and BÜHLMANN, P. (2021). Supplement to "Stabilizing variable selection and regression." https://doi.org/10.1214/21-AOAS1487SUPPA, https://doi.org/10.1214/21-AOAS1487SUPPB

RICHARDSON, T. and ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, Univ. Washington Series. Working Paper 128, 30 April 2013.

ROJAS-CARULLA, M., SCHÖLKOPF, B., TURNER, R. and PETERS, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res*. **19** Paper No. 36, 34. MR3862443

ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P. and PETERS, J. (2021). Anchor regression: Heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **83** 215–246. MR4250274 https://doi.org/10.1111/rssb.12398

ROUMELIOTIS, T. I., WILLIAMS, S. P., GONÇALVES, E., ALSINET, C., DEL CASTILLO VELASCO-HERRERA, M., ABEN, N., GHAVIDEL, F. Z., MICHAUT, M., SCHUBERT, M. et al. (2017). Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell Rep*. **20** 2201–2214.

ROY, S., SLEIMAN, M. B., JHA, P., WILLIAMS, E. G., INGELS, J. F., CHAPMAN, C. J., MCCARTY, M. S., HOOK, M., SUN, A. et al. (2019). Modulation of longevity by diet, and youthful body weight, but not by weight gain after maturity. Preprint bioRxiv:776559.

SCHÖLKOPF, B., JANZING, D., PETERS, J., SGOURITSA, E., ZHANG, K. and MOOIJ, J. M. (2012). On causal and anticausal learning. In *Proceedings of the* 29*th International Conference on Machine Learning* (*ICML*) 1255–1262. Omnipress.

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

SHAH, R. D. and BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 113–135. MR3744714 https://doi.org/10.1111/rssb.12234

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

WANG, S., NAN, B., ROSSET, S. and ZHU, J. (2011). Random Lasso. *Ann. Appl. Stat.* **5** 468–485. MR2810406 https://doi.org/10.1214/10-AOAS377

WILLIAMS, E. G., PFISTER, N., ROY, S., STATZER, S., INGELS, J., BOHL, C., HASSAN, M., ČUKLINA, J., BÜHLMANN, P. et al. (2020). Multi-omic profiling of the liver across diets and age in a diverse mouse population. Preprint bioRxiv. Available at https://www.biorxiv.org/content/10.1101/2020.08.20.222968v2.

WRIGHT, S. (1921). Correlation and causation. *J. Agric. Res.* **20** 557–585.

YU, B. (2013). Stability. *Bernoulli* **19** 1484–1500. MR3102560 https://doi.org/10.3150/13-BEJSP14

YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning* (*ICML*) 819–827.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1247–1272
https://doi.org/10.1214/21-AOAS1465

QINI-BASED UPLIFT REGRESSION

BY MOULOUD BELBAHRI[1,*], ALEJANDRO MURUA[1,†], OLIVIER GANDOUET[2] AND VAHID PARTOVI NIA[3]

[1]*Département de mathématiques et de statistique, Université de Montréal,* [*]*mouloud.belbahri@umontreal.ca;* [†]*alejandro.murua@umontreal.ca*

[2]*Advanced Analytics, Research and Development, TD Insurance, olivier.gandouet@tdassurance.com*

[3]*Department of Mathematics and Industrial Engineering, École Polytechnique de Montréal, vahid.partovinia@polymtl.ca*

Uplift models provide a solution to the problem of isolating the marketing effect of a campaign. For customer churn reduction, uplift models are used to identify the customers who are likely to respond positively to a retention activity, *only* if targeted, and to avoid wasting resources on customers that are very likely to switch to another company. In practice, the uplift models performance is measured by the Qini coefficient. We introduce a Qini-based uplift regression model to analyze a large insurance company's retention marketing campaign. Our approach is based on logistic regression models. We show that a Qini-optimized uplift model acts as a regularizing factor for uplift, much as a penalized likelihood model does for regression. This results in interpretable models with few relevant explanatory variables. Our results show that Qini-based parameter estimation significantly improves the Qini prediction performance of uplift models.

REFERENCES

ALEMI, F., ERDMAN, H., GRIVA, I. and EVANS, C. H. (2009). Improved statistical methods are needed to advance personalized medicine. *The Open Translational Medicine Journal* **1** 16.

BELBAHRI, M., GANDOUET, O., MURUA, A. and PARTOVI NIA, V. (2020). tools4uplift: Tools for Uplift Modeling. R package version 1.0.0.

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

COUSSEMENT, K., HARRIGAN, P. and BENOIT, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Syst. Appl.* **42** 8403–8412.

DEVRIENDT, F., MOLDOVAN, D. and VERBEKE, W. (2018). A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* **6** 13–41. https://doi.org/10.1089/big.2017.0104

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FRANK, L. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version 4.1.

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

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737 https://doi.org/10.1214/07-AOAS131

GINI, C. (1997). Concentration and dependency ratios. *Riv. Polit. Econ.* **87** 769–792.

GUELMAN, L. (2014). uplift: Uplift Modeling. R package version 0.3.5.

Key words and phrases. Casual inference, Kendall's correlation, lasso, logistic regression, marketing, derivative-free optimization.

GUELMAN, L., GUILLÉN, M. and PÉREZ-MARÍN, A. M. (2012). Random forests for uplift modeling: An insurance customer retention case. In *Modeling and Simulation in Engineering, Economics and Management* 123–133. Springer, Berlin.

GUTIERREZ, P. and GÉRARDY, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs* 1–13.

GUYON, I. and ELISSEEFF, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** 1157–1182.

HAND, D. J. and YU, K. (2001). Idiot's Bayes not so stupid after all. *Int. Stat. Rev.* **69** 385–398.

HANSOTIA, B. J. and RUKSTALES, B. (2001). Direct marketing for multichannel retailers: Issues, challenges and solutions. *J. Database Mark. Cust. Strategy Manag.* **9** 259–266.

HANSOTIA, B. and RUKSTALES, B. (2002). Incremental value modeling. *J. Interact. Mark.* **16** 35.

HANSSENS, D. M., PARSONS, L. J. and SCHULTZ, R. L. (2003). *Market Response Models: Econometric and Time Series Analysis* **12**. Springer, Berlin.

HASTIE, T., TAYLOR, J., TIBSHIRANI, R. and WALTHER, G. (2007). Forward stagewise regression and the monotone lasso. *Electron. J. Stat.* **1** 1–29. MR2312144 https://doi.org/10.1214/07-EJS004

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

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

KANE, K., LO, V. S. and ZHENG, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *J. Mark. Anal.* **2** 218–238.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.

LO, V. S. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explor. Newsl.* **4** 78–86.

LORENZ, M. O. (1905). Methods of measuring the concentration of wealth. *Publ. Amer. Stat. Assoc.* **9** 209–219.

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245. MR0533252 https://doi.org/10.2307/1268522

MONTGOMERY, D. C., PECK, E. A. and VINING, G. G. (2012). *Introduction to Linear Regression Analysis* **821**. Wiley, New York.

NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. MR3363409 https://doi.org/10.1093/comjnl/7.4.308

NEYMAN (1923). On the application of probability theory to agricultural experiments. *Ann. Agric. Sci.*.

PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291 https://doi.org/10.1214/09-SS057

RADCLIFFE, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Market J. Direct Market Assoc. Anal. Council* **1** 14–21.

RADCLIFFE, N. and SURRY, P. (1999). Differential response analysis: Modeling true response by isolating the effect of a single action. Credit Scoring and Credit Control VI. Edinburgh, Scotland.

RADCLIFFE, N. J. and SURRY, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RZEPAKOWSKI, P. and JAROSZEWICZ, S. (2010). Decision trees for uplift modeling. In 2010 *IEEE International Conference on Data Mining* 441–450. IEEE Comput. Soc., Los Alamitos.

RZEPAKOWSKI, P. and JAROSZEWICZ, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowl. Inf. Syst.* **32** 303–327.

SMITH, R. E. and SWINYARD, W. R. (1982). Information response models: An integrated approach. *J. Mark.* **46** 81–93.

SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10** 141–158.

SU, X., KANG, J., FAN, J., LEVINE, R. A. and YAN, X. (2012). Facilitating score and causal inference trees for large observational studies. *J. Mach. Learn. Res.* **13** 2955–2994. MR2997717

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

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1273–1290
https://doi.org/10.1214/21-AOAS1462

ORTHOGONAL SUBSAMPLING FOR BIG DATA LINEAR REGRESSION

BY LIN WANG[1], JAKE ELMSTEDT[2,*], WENG KEE WONG[3] AND HONGQUAN XU[2,†]

[1]*Department of Statistics, George Washington University, linwang@gwu.edu*
[2]*Department of Statistics, University of California, Los Angeles, *jake.r.kramer@gmail.com; †hqxu@stat.ucla.edu*
[3]*Department of Biostatistics, University of California, Los Angeles, wkwong@ucla.edu*

The dramatic growth of big datasets presents a new challenge to data storage and analysis. Data reduction, or subsampling, that extracts useful information from datasets is a crucial step in big-data analysis. We propose an orthogonal subsampling (OSS) approach for big data with a focus on linear regression models. The approach is inspired by the fact that an orthogonal array of two levels provides the best experimental design for linear regression models in the sense that it minimizes the average variance of the estimated parameters and provides the best predictions. The merits of OSS are three-fold: (i) it is easy to implement and fast; (ii) it is suitable for distributed parallel computing and ensures the subsamples selected in different batches have no common data points, and (iii) it outperforms existing methods in minimizing the mean squared errors of the estimated parameters and maximizing the efficiencies of the selected subsamples. Theoretical results and extensive numerical results show that the OSS approach is superior to existing subsampling approaches. It is also more robust to the presence of interactions among covariates, and, when they do exist, OSS provides more precise estimates of the interaction effects than existing methods. The advantages of OSS are also illustrated through analysis of real data.

REFERENCES

ATKINSON, A. C., DONEV, A. N. and TOBIAS, R. D. (2007). *Optimum Experimental Designs, with SAS*. *Oxford Statistical Science Series* **34**. Oxford Univ. Press, Oxford. MR2323647

BOX, G. E. P. and WILSON, K. B. (1951). On the experimental attainment of optimum conditions. *J. Roy. Statist. Soc. Ser. B* **13** 1–38; discussion: 38–45. MR0046009

BUZA, K. (2014). Feedback prediction for blogs. In *Data Analysis, Machine Learning and Knowledge Discovery* 145–152. Springer, Berlin.

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

CHALONER, K. and LARNTZ, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Plann. Inference* **21** 191–208. MR0985457 https://doi.org/10.1016/0378-3758(89)90004-9

CHENG, C.-S. (1980). Orthogonal arrays with variable numbers of symbols. *Ann. Statist.* **8** 447–453. MR0560740

CHENG, C.-S. (1997). $E(s^2)$-optimal supersaturated designs. *Statist. Sinica* **7** 929–939. MR1488651

DEY, A. and MUKERJEE, R. (1999). *Fractional Factorial Plans. Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1679441 https://doi.org/10.1002/9780470316986

DRINEAS, P., MAGDON-ISMAIL, M., MAHONEY, M. W. and WOODRUFF, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.* **13** 3475–3506. MR3033372

DUA, D. and GRAFF, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

FAN, Y. and SUN, J. (2020). Subsampling Winner Algorithm for Feature Selection in Large Regression Data. Preprint. Available at arXiv:2002.02903.

FANG, K.-T., LI, R. and SUDJIANTO, A. (2006). *Design and Modeling for Computer Experiments*. *Chapman & Hall/CRC Computer Science and Data Analysis Series*. CRC Press/CRC, Boca Raton, FL. MR2223960

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

HEDAYAT, A. S., SLOANE, N. J. A. and STUFKEN, J. (1999). *Orthogonal Arrays*: *Theory and Applications*. *Springer Series in Statistics*. Springer, New York. MR1693498 https://doi.org/10.1007/978-1-4612-1478-6

Key words and phrases. Data reduction, experimental design, optimal design, orthogonal array.

KIEFER, J. (1959). Optimum experimental designs. *J. Roy. Statist. Soc. Ser. B* **21** 272–319. MR0113263
LI, W. W. and WU, C. F. J. (1997). Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics* **39** 171–179. MR1452345 https://doi.org/10.2307/1270905
MA, P. and SUN, X. (2015). Leveraging for big data regression. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 70–76. MR3348722 https://doi.org/10.1002/wics.1324
MAJUMDAR, J., NARASEEYAPPA, S. and ANKALAKI, S. (2017). Analysis of agriculture data using data mining techniques: Application of big data. *J. Big Data* **4** 20.
MELIE-GARCIA, L., DRAGANSKI, B., ASHBURNER, J. and KHERIF, F. (2018). Multiple linear regression: Bayesian inference for distributed and Big Data in the Medical Informatics Platform of the Human Brain Project. BioRxiv https://doi.org/10.1101/242883
MILLER, A. J. and NGUYEN, N.-K. (1994). Algorithm AS 295: A Fedorov exchange algorithm for D-optimal design. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **43** 669–677.
NGUYEN, N.-K. and MILLER, A. J. (1992). A review of some exchange algorithms for constructing discrete *D*-optimal designs. *Comput. Statist. Data Anal.* **14** 489–498. MR1192218 https://doi.org/10.1016/0167-9473(92)90064-M
OVERSTALL, A. M. and WOODS, D. C. (2017). Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59** 458–470. MR3740963 https://doi.org/10.1080/00401706.2016.1251495
ROYLE, J. A. (2002). Exchange algorithms for constructing large spatial designs *J. Statist. Plann. Inference* **100** 121–134. MR1877182 https://doi.org/10.1016/S0378-3758(01)00127-6
SABER, A. Y. and ALAM, A. R. (2017). Short term load forecasting using multiple linear regression for big data. In 2017 *IEEE Symposium Series on Computational Intelligence* (*SSCI*) 1–6. IEEE, New York.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
WANG, H. (2019). Divide-and-conquer information-based optimal subdata selection algorithm. *J. Stat. Theory Pract.* **13** Paper No. 46, 19. MR3978445 https://doi.org/10.1007/s42519-019-0048-5
WANG, H., YANG, M. and STUFKEN, J. (2019). Information-based optimal subdata selection for big data linear regression. *J. Amer. Statist. Assoc.* **114** 393–405. MR3941263 https://doi.org/10.1080/01621459.2017.1408468
WANG, Y., YU, A. W. and SINGH, A. (2017). On computationally tractable selection of experiments in measurement-constrained regression models. *J. Mach. Learn. Res.* **18** Paper No. 143, 41. MR3763777
WANG, L., ELMSTEDT, J., WONG, W. K. and XU, H. (2021). Supplement to "Orthogonal subsampling for big data linear regression." https://doi.org/10.1214/21-AOAS1462SUPP
WU, C. F. J. (1993). Construction of supersaturated designs through partially aliased interactions. *Biometrika* **80** 661–669. MR1248029 https://doi.org/10.1093/biomet/80.3.661
XU, H. (2002). An algorithm for constructing orthogonal and nearly-orthogonal arrays with mixed levels and small runs. *Technometrics* **44** 356–368. MR1939683 https://doi.org/10.1198/004017002188618554
XU, H. (2003). Minimum moment aberration for nonregular designs and supersaturated designs. *Statist. Sinica* **13** 691–708. MR1997169
XU, H., PHOA, F. K. H. and WONG, W. K. (2009). Recent developments in nonregular fractional factorial designs. *Stat. Surv.* **3** 18–46. MR2520978 https://doi.org/10.1214/08-SS040
ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1291–1307
https://doi.org/10.1214/21-AOAS1446

ESTROGEN RECEPTOR EXPRESSION ON BREAST CANCER PATIENTS' SURVIVAL UNDER SHAPE-RESTRICTED COX REGRESSION MODEL

BY JING QIN[1], GENG DENG[2], JING NING[3,*], AO YUAN[4] AND YU SHEN[3,†]

[1]*National Institution of Allergy and Infectious Diseases, jingqin@niaid.nih.gov*
[2]*Wells Fargo, gengdeng@gmail.com*
[3]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, *jning@mdanderson.org; †yshen@mdanderson.org*
[4]*Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, AY312@georgetown.edu*

For certain subtypes of breast cancer, study findings show that their level of estrogen receptor expression is associated with their risk of cancer death and also suggest a nonlinear effect on the hazard of death. A flexible form of the proportional hazards model, $\lambda(t|x,\mathbf{z}) = \lambda(t)\exp(\mathbf{z}^T\boldsymbol{\beta})q(x)$, is desirable to facilitate a rich class of covariate effect on a survival outcome to provide meaningful insight, where the functional form of $q(x)$ is not specified except for its shape. Prior biologic knowledge on the shape of the underlying distribution of the covariate effect in regression models can be used to enhance statistical inference. Despite recent progress, major challenges remain for the semiparametric shape-restricted inference due to lack of practical and efficient computational algorithms to accomplish nonconvex optimization. We propose an alternative algorithm to maximize the full log-likelihood with two sets of parameters iteratively under monotone constraints. The first set consists of the nonparametric estimation of the monotone-restricted function $q(x)$, while the second set includes estimating the baseline hazard function and other covariate coefficients. The iterative algorithm, in conjunction with the pool-adjacent-violators algorithm, makes the computation efficient and practical. The jackknife resampling effectively reduces the estimator bias, when sample size is small. Simulations show that the proposed method can accurately capture the underlying shape of $q(x)$ and outperforms the estimators when $q(x)$ in the Cox model is misspecified. We apply the method to model the effect of estrogen receptor on breast cancer patients' survival.

REFERENCES

ANCUKIEWICZ, M., FINKELSTEIN, D. M. and SCHOENFELD, D. A. (2003). Modelling the relationship between continuous covariates and clinical events using isotonic regression. *Stat. Med.* **22** 3151–3159.

BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA. MR3444832

BRESLOW, N. E. (1972). Discussion of the paper by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 216–217.

CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348 https://doi.org/10.1111/rssb.12137

CHEN, K., GUO, S., SUN, L. and WANG, J.-L. (2010). Global partial likelihood for nonparametric proportional hazards models. *J. Amer. Statist. Assoc.* **105** 750–760. MR2724858 https://doi.org/10.1198/jasa.2010.tm08636

CHUNG, Y., IVANOVA, A., HUDGENS, M. G. and FINE, J. P. (2018). Partial likelihood estimation of isotonic proportional hazards models. *Biometrika* **105** 133–148. MR3768870 https://doi.org/10.1093/biomet/asx064

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237 https://doi.org/10.1111/j.1467-9868.2010.00753.x

Key words and phrases. Concave or convex function, Cox proportional hazards model, jackknife bias correction, pool adjacent violators algorithm, shape-restricted inference.

DOSS, C. R. and WELLNER, J. A. (2016). Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Statist.* **44** 954–981. MR3485950 https://doi.org/10.1214/15-AOS1394

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA. MR0659849

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

FUJII, T., KOGAWA, T., DONG, W., SAHIN, A. A., MOULDER, S., LITTON, J. K., TRIPATHY, D., IWAMOTO, T., HUNT, K. K. et al. (2017). Revisiting the definition of estrogen receptor positivity in HER2-negative primary breast cancer. *Ann. Oncol.* **28** 2420–2428.

GORSKI, J., PFEUFFER, F. and KLAMROTH, K. (2007). Biconvex sets and optimization with biconvex functions: A survey and extensions. *Math. Methods Oper. Res.* **66** 373–407. MR2357657 https://doi.org/10.1007/s00186-007-0161-1

GRAMBSCH, P. M. and THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81** 515–526. MR1311094 https://doi.org/10.1093/biomet/81.3.515

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

IWAMOTO, T., BOOSER, D., VALERO, V., MURRAY, J. L., KOENIG, K., ESTEVA, F. J., UENO, N. T., ZHANG, J., SHI, W. et al. (2012). Estrogen receptor (ER) mRNA and ER-related gene expression in breast cancers that are 1% to 10% ER-positive by immunohistochemistry. *J. Clin. Oncol.* **30** 729–734.

LANCASTER, T. (2000). The incidental parameter problem since 1948. *J. Econometrics* **95** 391–413. MR1752336 https://doi.org/10.1016/S0304-4076(99)00044-5

MIRATRIX, L. W., WAGER, S. and ZUBIZARRETA, J. R. (2018). Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika* **105** 103–114. MR3768868 https://doi.org/10.1093/biomet/asx077

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–465. MR1803168 https://doi.org/10.2307/2669386

MURRAY, T. A., HOBBS, B. P., SARGENT, D. J. and CARLIN, B. P. (2016). Flexible Bayesian survival modeling with semiparametric time-dependent and shape-restricted covariate effects. *Bayesian Anal.* **11** 381–402. MR3471995 https://doi.org/10.1214/15-BA954

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113 https://doi.org/10.2307/1914288

NOCEDAL, J. and WRIGHT, S. J. (2006). *Nonlinear Equations*. Springer, Berlin.

QIN, J., DENG, G., NING, J., YUAN, A. and SHEN, Y. (2021). Supplement to "Estrogen receptor expression on breast cancer patients' survival under shape-restricted Cox regression model." https://doi.org/10.1214/21-AOAS1446SUPP

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360. MR0081040 https://doi.org/10.1093/biomet/43.3-4.353

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

SHAO, J. et al. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17** 1176–1197. MR1015145 https://doi.org/10.1214/aos/1176347263

SHAO, J. and TU, D. S. (1995). *The Jackknife and Bootstrap. Springer Series in Statistics*. Springer, New York. MR1351010 https://doi.org/10.1007/978-1-4612-0795-5

TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82** 559–567. MR0898359

YI, M., HUO, L., KOENIG, K. B., MITTENDORF, E. A., MERIC-BERNSTAM, F., KUERER, H. M., BEDROSIAN, I., BUZDAR, A. U., SYMMANS, W. F. et al. (2014). Which threshold for ER positivity? A retrospective study based on 9639 patients. *Ann. Oncol.* **25** 1004–1011.

ZHOU, M. (2016). *Empirical Likelihood Method in Survival Analysis. Chapman & Hall/CRC Biostatistics Series*. CRC Press, Boca Raton, FL. MR3616660

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1308–1328
https://doi.org/10.1214/21-AOAS1445

MODELING PAST EVENT FEEDBACK THROUGH BIOMARKER DYNAMICS IN THE MULTISTATE EVENT ANALYSIS FOR CARDIOVASCULAR DISEASE DATA

BY CHUOXIN MA[1,*], HONGSHENG DAI[2] AND JIANXIN PAN[1,†]

[1]*Department of Mathematics, The University of Manchester,* [*]*chuoxin.ma@manchester.ac.uk;* [†]*Jianxin.Pan@manchester.ac.uk*
[2]*Department of Mathematical Sciences, University of Essex, hdaia@essex.ac.uk*

In cardiovascular studies we often observe ordered multiple events along disease progression which are, essentially, a series of recurrent events and terminal events with competing risk structure. One of the main interests is to explore the event specific association with the dynamics of longitudinal biomarkers. A new statistical challenge arises when the biomarkers carry information from the past event history, providing feedbacks for the occurrences of future events and, particularly, when these biomarkers are only intermittently observed with measurement errors. In this paper we propose a novel modeling framework where the recurrent events and terminal events are modeled as multistate processes and the longitudinal covariates that account for event feedbacks are described by random effects models. Considering the nature of long-term observation in cardiac studies, flexible models with semiparametric coefficients are adopted. To improve computation efficiency, we develop an one-step estimator of the regression coefficients and derive their asymptotic variances for the computation of the confidence intervals, based on the proposed asymptotically unbiased estimating equation. Simulation studies show that the naive estimators, which either ignore the past event feedbacks or the measurement errors, are biased. Our method achieves better coverage probability, compared to the naive methods. The model is motivated and applied to a dataset from the Atherosclerosis Risk in Communities Study.

REFERENCES

AALEN, O. O., BORGAN, Ø. and GJESSING, H. K. (2008). *Survival and Event History Analysis*: *A Process Point of View*. *Statistics for Biology and Health*. Springer, New York. MR2449233 https://doi.org/10.1007/978-0-387-68560-1

AALEN, O. O., FOSEN, J., WEEDON-FEKJÆR, H., BORGAN, Ø. and HUSEBYE, E. (2004). Dynamic analysis of multivariate failure time data. *Biometrics* **60** 764–773. MR2089453 https://doi.org/10.1111/j.0006-341X.2004.00227.x

BARRETT, J. K., HUILLE, R., PARKER, R., YANO, Y. and GRISWOLD, M. (2019). Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC study. *Stat. Med*. **38** 1855–1868. MR3934823 https://doi.org/10.1002/sim.8074

BEDAIR, K., HONG, Y., LI, J. and AL-KHALIDI, H. R. (2016). Multivariate frailty models for multi-type recurrent event data and its application to cancer prevention trial. *Comput. Statist. Data Anal*. **101** 161–173. MR3504843 https://doi.org/10.1016/j.csda.2016.01.018

BEYERSMANN, J., ALLIGNOL, A. and SCHUMACHER, M. (2012). *Competing Risks and Multistate Models with R*. *Use R!* Springer, New York. MR3025354 https://doi.org/10.1007/978-1-4614-2035-4

BORGAN, Ø., FIACCONE, R. L., HENDERSON, R. and BARRETO, M. L. (2007). Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil. *Scand. J. Stat*. **34** 53–69. MR2325242 https://doi.org/10.1111/j.1467-9469.2006.00525.x

CAI, Z. and SUN, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand. J. Stat*. **30** 93–111. MR1963895 https://doi.org/10.1111/1467-9469.00320

Key words and phrases. Asymptotically unbiased estimating equation, cardiovascular disease, measurement errors, multistate models, ordered multiple event, past event feedback, semiparametric coefficients.

CAO, J. and YAO, W. (2012). Semiparametric mixture of binomial regression with a degenerate component. *Statist. Sinica* **22** 27–46. MR2933166 https://doi.org/10.5705/ss.2010.062

CAO, H., CHURPEK, M. M., ZENG, D. and FINE, J. P. (2015). Analysis of the proportional hazards model with sparse longitudinal covariates. *J. Amer. Statist. Assoc.* **110** 1187–1196. MR3420694 https://doi.org/10.1080/01621459.2014.957289

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842 https://doi.org/10.2307/2965697

COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events. Statistics for Biology and Health*. Springer, New York. MR3822124

COOK, R. J., YI, G. Y., LEE, K.-A. and GLADMAN, D. D. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics* **60** 436–443. MR2066278 https://doi.org/10.1111/j.0006-341X.2004.00188.x

DAI, H. and PAN, J. (2018). Joint modelling of survival and longitudinal data with informative observation times. *Scand. J. Stat.* **45** 571–589. MR3858947 https://doi.org/10.1111/sjos.12314

DE LA SIERRA, A., SEGURA, J., GOROSTIDI, M., BANEGAS, J. R., DE LA CRUZ, J. J. and RUILOPE, L. M. (2010). Diurnal blood pressure variation, risk categories and antihypertensive treatment. *Hypertens. Res.* **33** 767–771. https://doi.org/10.1038/hr.2010.111

ELISAF, M. S., KALAITZIDIS, R. G., GOUDEVENOS, J. A., KATSARAKI, A. E., SIDERIS, D. A. and SIAMOPOULOS, K. C. (1999). Blood pressure profile in patients with microvascular angina. *Coron. Artery Dis.* **10** 257–259.

FISHER, L. D. and LIN, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* **20** 145–157.

GJESSING, H. K., RØYSLAND, K., PENA, E. A. and AALEN, O. O. (2010). Recurrent events and the exploding Cox model. *Lifetime Data Anal.* **16** 525–546. MR2726223 https://doi.org/10.1007/s10985-010-9180-y

HUANG, Y.-H., HWANG, W.-H. and CHEN, F.-Y. (2016). Improving efficiency using the Rao–Blackwell theorem in corrected and conditional score estimation methods for joint models. *Biometrics* **72** 1136–1144. MR3591598 https://doi.org/10.1111/biom.12510

HUANG, M. and YAO, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Amer. Statist. Assoc.* **107** 711–724. MR2980079 https://doi.org/10.1080/01621459.2012.682541

IP, E. H., EFENDI, A., MOLENBERGHS, G. and BERTONI, A. G. (2015). Comparison of risks of cardiovascular events in the elderly using standard survival analysis and multiple-events and recurrent-events methods. *BMC Med. Res. Methodol.* **15** 15. https://doi.org/10.1186/s12874-015-0004-3

KIM, S., ZENG, D., CHAMBLESS, L. and LI, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Stat. Biosci.* **4** 262–281.

MA, C., DAI, H. and PAN, J. (2021). Supplement to "Modeling past event feedback through biomarker dynamics in the multistate event analysis for cardiovascular disease data." https://doi.org/10.1214/21-AOAS1445SUPPA, https://doi.org/10.1214/21-AOAS1445SUPPB

MILOSLAVSKY, M., KELEŞ, S., VAN DER LAAN, M. J. and BUTLER, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 239–257. MR2035769 https://doi.org/10.1111/j.1467-9868.2004.00442.x

PEÑA, E. A. (2006). Dynamic modelling and statistical analysis of event times. *Statist. Sci.* **21** 487–500. MR2369983 https://doi.org/10.1214/088342306000000349

ROGERS, J. K., YAROSHINSKY, A., POCOCK, S. J., STOKAR, D. and POGODA, J. (2016). Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. *Stat. Med.* **35** 2195–2205. MR3513508 https://doi.org/10.1002/sim.6853

SONG, X. and WANG, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics* **64** 557–566. MR2432426 https://doi.org/10.1111/j.1541-0420.2007.00890.x

SONG, X. and WANG, L. (2017). Partially time-varying coefficient proportional hazards models with error-prone time-dependent covariates—an application to the AIDS clinical trial group 175 data. *Ann. Appl. Stat.* **11** 274–296. MR3634324 https://doi.org/10.1214/16-AOAS1003

TSIATIS, A. A. and DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. MR1844844 https://doi.org/10.1093/biomet/88.2.447

WANG, C. Y. (2006). Corrected score estimator for joint modeling of longitudinal and failure time data. *Statist. Sinica* **16** 235–253. MR2256090

WATTANAKIT, K., FOLSOM, A. R., CHAMBLESS, L. E. and NIETO, F. J. (2005). Risk factors for cardiovascular event recurrence in the atherosclerosis risk in communities (ARIC) study. *Am. Heart J.* **149** 606–612.

XIAO, W., LU, W. and ZHANG, H. H. (2016). Joint structure selection and estimation in the time-varying coefficient Cox model. *Statist. Sinica* **26** 547–567. MR3497759

YAN, J. and HUANG, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics* **68** 419–428. MR2959608 https://doi.org/10.1111/j.1541-0420.2011.01692.x

YANG, W., JEPSON, C., XIE, D., ROY, J. A., SHOU, H., HSU, J. Y., ANDERSON, A. H., LANDIS, J. R., HE, J. et al. (2017). Statistical methods for recurrent event analysis in cohort studies of CKD. *Clin. J. Amer. Soc. Nephrol*. **12** 2066–2073.

ZHAO, Y., PARK, J., IZADNEGAHDAR, M., LEE, M., KHAN, N., RABKIN, S., GUAN, M., GRUBISIC, M., PENG, D. et al. (2017). Factors associated with very low diastolic blood pressure in the sprint trial. *Can. J. Cardiol*. **33** S199.

ZHOU, J., ZHANG, J., MCLAIN, A. C., LU, W., SUI, X. and HARDIN, J. W. (2019). A varying-coefficient generalized odds rate model with time-varying exposure: An application to fitness and cardiovascular disease mortality. *Biometrics* **75** 853–863. MR4012091 https://doi.org/10.1111/biom.13057

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1329–1342
https://doi.org/10.1214/20-AOAS1415

A MULTIVARIATE SPATIOTEMPORAL CHANGE-POINT MODEL OF OPIOID OVERDOSE DEATHS IN OHIO

BY STACI A. HEPLER, LANCE A. WALLER AND DAVID M. KLINE

[1]*Department of Mathematics and Statistics, Wake Forest University, heplersa@wfu.edu*
[2]*Department of Biostatistics and Bioinformatics, Emory University, lwaller@emory.edu*
[3]*Center for Biostatistics, Department of Biomedical Informatics, Ohio State University, david.kline@osumc.edu*

Ohio is one of the states most impacted by the opioid epidemic and experienced the second highest age-adjusted fatal drug overdose rate in 2017. Initially it was believed prescription opioids were driving the opioid crisis in Ohio. However, as the epidemic evolved, opioid overdose deaths due to fentanyl have drastically increased. In this work we develop a Bayesian multivariate spatiotemporal model for Ohio county overdose death rates from 2007 to 2018 due to different types of opioids. The log-odds are assumed to follow a spatially varying change point regression model. By assuming the regression coefficients are a multivariate conditional autoregressive process, we capture spatial dependence within each drug type and also dependence across drug types. The proposed model allows us to not only study spatiotemporal trends in overdose death rates but also to detect county-level shifts in these trends over time for various types of opioids.

REFERENCES

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.

BERCHUCK, S. I., MWANZA, J.-C. and WARREN, J. L. (2019). A spatially varying change points model for monitoring glaucoma progression using visual field data. *Spat. Stat.* **30** 1–26. MR3921304 https://doi.org/10.1016/j.spasta.2019.02.001

BERLINER, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods. Fund. Theories Phys.* **79** 15–22. Kluwer Academic, Dordrecht. MR1446713

CARLIN, B. P. and BANERJEE, S. (2003). Hierarchical multivariate CAR models for Spatio-temporally correlated survival data. *Bayesian Statistics* **7** 45–63.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2012). HIV infection and HIV-associated behaviors among injecting drug users—20 cities, United States, 2009. *Morb. Mort. Wkly. Rep.* 61 133–138.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2019). National survey on drug use and health, 2013 and 2014. Internet. Available at https://www.cdc.gov/opioids/index.html.

CICCARONE, D. (2019). The triple wave epidemic: Supply and demand drivers of the US opioid overdose crisis. *Int. J. Drug Policy* **71** 183–188. https://doi.org/10.1016/j.drugpo.2019.01.010

COMPTON, W. M., JONES, C. M. and BALDWIN, G. T. (2016). Relationship between nonmedical prescription-opioid use and heroin use. *N. Engl. J. Med.* **374** 154–163.

DANIULAITYTE, R., JUHASCIK, M. P., STRAYER, K. E., SIZEMORE, I. E., HARSHBARGER, K. E., ANTONIDES, H. M. and CARLSON, R. R. (2017). Overdose deaths related to fentanyl and its analogs—Ohio, January-February 2017. *Morb. Mort. Wkly. Rep.* **66** 904–908. https://doi.org/10.15585/mmwr.mm6634a3

DART, R. C., SURRATT, H. L., CICERO, T. J., PARRINO, M. W., SEVERTSON, S. G., BUCHER-BARTELSON, B. and GREEN, J. L. (2015). Trends in opioid analgesic abuse and mortality in the United States. *N. Engl. J. Med.* **372** 241–248. https://doi.org/10.1056/NEJMsa1406143

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 https://doi.org/10.1080/10618600.2016.1172487

DEGENHARDT, L., LARNEY, S., KIMBER, J., FARRELL, M. and HALL, W. (2015). Excess mortality among opioid-using patients treated with oral naltrexone in Australia. *Drug Alcohol Rev.* **34** 90–96. https://doi.org/10.1111/dar.12205

Key words and phrases. Multivariate conditional autoregressive, change point, Bayesian, spatial rates.

DELCHER, C., WAGENAAR, A. C., GOLDBERGER, B. A., COOK, R. L. and MALDONADO-MOLINA, M. M. (2015). Abrupt decline in oxycodone-caused mortality after implementation of Florida's prescription drug monitoring program. *Drug Alcohol Depend*. **150** 63–68.

DIVISION OF UNINTENTIONAL INJURY PREVENTION (2017). Drug overdose data. Internet. Available at https://www.cdc.gov/drugoverdose/data/statedeaths.html.

DRUG ENFORCEMENT AGENCY (2016). 2016 national drug threat assessment. Internet. Available at https://www.dea.gov/resource-center/2016%20NDTA%20Summary.pdf.

EVANS, E., LI, L., MIN, J., HUANG, D., URADA, D., LIU, L., HSER, Y. I. and NOSYK, B. (2015). Mortality among individuals accessing pharmacological treatment for opioid dependence in California, 2006-10. *Addiction* **110** 996–1005.

GELFAND, A. E. and VOUNATSOU, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4** 11–25. https://doi.org/10.1093/biostatistics/4.1.11

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

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

GOVERNOR'S CABINET OPIATE ACTION TEAM (2012). Attacking Ohio's opiate epidemic. Online. Available at https://mha.ohio.gov/Researchers-and-Media/Combating-the-Opioid-Crisis; accessed 6-September-2017.

GOVERNOR'S CABINET OPIATE ACTION TEAM (2018). Combating the opiate crisis in Ohio. Online. Available at https://mha.ohio.gov/Portals/0/assets/ResearchersAndMedia/Combating%20Opiate%20Abuse/Combating-the-Opiate-Crisis_SEPT-2018.pdf?ver=2018-11-29-113014-833; accessed 25-February-2020.

HEPLER, S. A., WALLER, L. A. and KLINE, D. M. (2021a). Supplement to "A multivariate spatio temporal change-point model of opioid overdose deaths in Ohio." https://doi.org/10.1214/20-AOAS1415SUPPA.

HEPLER, S. A., WALLER, L. A. and KLINE, D. M. (2021b). Data Files and Code for "A multivariate spatio-temporal change point model of opioid overdose deaths in Ohio" https://doi.org/10.1214/20-AOAS1415SUPPB.

HSER, Y. I., SAXON, A. J., HUANG, D., HASSON, A., THOMAS, C., HILLHOUSE, M., JACOBS, P., TERUYA, C., MCLAUGHLIN, P. et al. (2014). Treatment retention among patients randomized to buprenorphine/naloxone compared to methadone in a multi-site trial. *Addiction* **109** 79–87.

JONES, C. M. (2013). Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers—United States, 2002-2004 and 2008-2010. *Drug Alcohol Depend.* **132** 95–100.

LAWSON, A. B. (2020). NIMBLE for Bayesian disease mapping. *Spat. Spatio-Tempor. Epidemiol.* **33**.

MARS, S. G., ROSENBLUM, D. and CICCARONE, D. (2019). Illicit fentanyls in the opioid street market: Desired or imposed? *Addiction* **114** 774–780. https://doi.org/10.1111/add.14474

MARTINEZ-BENEITO, M. A., BOTELLA-ROCAMORA, P. and BANERJEE, S. (2017). Towards a multidimensional approach to Bayesian disease mapping. *Bayesian Anal.* **12** 239–259. MR3597574 https://doi.org/10.1214/16-BA995

OFFICE OF NATIONAL DRUG CONTROL POLICY EXECUTIVE, OFFICE OF THE PRESIDENT OF THE UNITED STATES (2011). Epidemic: Responding to America's prescription drug abuse crisis. Internet. Available at https://www.hsdl.org/?view&did=4609.

OHIO PUBLIC HEALTH DATA WAREHOUSE (2020). Ohio resident mortality data. Available at http://publicapps.odh.ohio.gov/EDW/DataCatalog; accessed February 2, 2020.

PACIOREK, C. (2009). Technical Vignette 5: Understanding intrinsic Gaussian Markov random field spatial models, including intrinsic conditional autoregressive models. Technical Report. Dept. Statistics, Univ. California, Berkeley and Dept. Biostatistics, Harvard School of Public Health.

PENM, J., MACKINNON, N. J., BOONE, J. M., CIACCIA, A., MCNAMEE, C. and WINSTANLEY, E. L. (2017). Strategies and policies to address the opioid epidemic: A case study of Ohio. *J. Am. Pharm. Assoc.* **57** S148–S153. https://doi.org/10.1016/j.japh.2017.01.001

PIANNTADOSI, S., BYAR, D. P. and GREEN, S. B. (1988). The ecological fallacy. *Am. J. Epidemiol.* **127** 893–904.

REMBERT, M., BETZ, M., FENG, B. and PARTRIDGE, M. (2017). Taking measure of Ohio's opioid crisis. Available at https://aede.osu.edu/sites/aede/files/publication_files/Swank%20-%20Taking%20Measure%20of%20Ohios%20Opioid%20Crisis.pdf.

RIGG, K. K., MARCH, S. J. and INCIARDI, J. A. (2010). Prescription drug abuse & diversion: Role of the pain clinic. *J. Drug Issues* **40** 681–702. https://doi.org/10.1177/002204261004000307

RUDD, R. A., ALESHIRE, N., ZIBBELL, J. E. and GLADDEN, R. M. (2016). Increases in drug and opioid overdose deaths–United States, 2000-2014. *Morb. Mort. Wkly. Rep.* **64** 1378–1382. https://doi.org/10.15585/mmwr.mm6450a3

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 https://doi.org/10.1201/9780203492024

SAMHSA (2017). Key substance use and mental health indicators in the United States: Results from the 2016 National Survey on Drug Use and Health. HHS Publication No. SMA 17-5044, NSDUH Series H-52. Available at http://www.samhsa.gov/.

SETH, P., RUDD, R. A., NOONAN, R. K. and HAEGERICH, T. M. (2018). Quantifying the epidemic of prescription opioid overdose deaths. *Am. J. Publ. Health* **108** 500–502. https://doi.org/10.2105/AJPH.2017.304265

SLAVOVA, S., O'BRIEN, D. B., CREPPAGE, K., DAO, D., FONDARIO, A., HAILE, E., HUME, B., LARGO, T. W., NGUYEN, C. et al. (2015). Drug overdose deaths: Let's get specific. *Public Health Reports* **130**.

TEESSON, M., MAREL, C., DARKE, S., ROSS, J., SLADE, T., BURNS, L., LYNSKEY, M., MEMEDOVIC, S., WHITE, J. et al. (2015). Long-term mortality, remission, criminality and psychiatric comorbidity of heroin dependence: 11-year findings from the Australian treatment outcome study. *Addiction* **110** 986–993. https://doi.org/10.1111/add.12860

WARREN, J. L., PINGALI, S. C. and WEINBERGER, D. M. (2017). Spatial variability in the persistence of pneumococcal conjugate vaccine-targeted pneumococcal serotypes among adults. *Epidemiology* **28** 119–126. https://doi.org/10.1097/EDE.0000000000000551

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194

WINSTANLEY, E. L., ZHANG, Y., MASHNI, R., SCHNEE, S., PENM, J., BOONE, J., MCNAMEE, C. and MACKINNON, N. J. (2018). Mandatory review of a prescription drug monitoring program and impact on opioid and benzodiazepine dispensing. *Drug Alcohol Depend.* **188** 169–174.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1343–1365
https://doi.org/10.1214/20-AOAS1429

TWO-STAGE CIRCULAR-CIRCULAR REGRESSION WITH ZERO INFLATION: APPLICATION TO MEDICAL SCIENCES

BY JAYANT JHA[1] AND PRAJAMITRA BHUYAN[2]

[1]*Institut de Neurosciences des Systèmes, Aix-Marseille University, jayantjha@gmail.com*
[2]*Department of Mathematics, Imperial College London, prajamitra.bhuyan@gmail.com*

This paper considers the modeling of zero-inflated circular measurements concerning real case studies from medical sciences. Circular-circular regression models have been discussed in the statistical literature and illustrated with various real-life applications. However, there are no models to deal with zero-inflated response as well as a covariate simultaneously. The Möbius transformation based two-stage circular-circular regression model is proposed, and the Bayesian estimation of the model parameters is suggested using the MCMC algorithm. Simulation results show the superiority of the performance of the proposed method over the existing competitors. The method is applied to analyse real datasets on astigmatism due to cataract surgery and abnormal gait related to orthopaedic impairment. The methodology proposed can assist in efficient decision making during treatment or postoperative care.

REFERENCES

BAKSHI, P. (2010). Evaluation of various surgical techniques in Brunescent cataracts. Unpublished thesis, Disha Eye Hospital, India.

BHATTACHARYA, S. and SENGUPTA, A. (2009). Bayesian analysis of semiparametric linear-circular models. *J. Agric. Biol. Environ. Stat.* **14** 33–65. MR2649679 https://doi.org/10.1198/jabes.2009.0003

BHUYAN, P., BISWAS, J., GHOSH, P. and DAS, K. (2019). A Bayesian two-stage regression approach of analysing longitudinal outcomes with endogeneity and incompleteness. *Stat. Model.* **19** 157–173. MR3921328 https://doi.org/10.1177/1471082X17747806

BISWAS, A., JHA, J. and DUTTA, S. (2016). Modelling circular random variables with a spike at zero. *Statist. Probab. Lett.* **109** 194–201. MR3434978 https://doi.org/10.1016/j.spl.2015.11.022

CAMERON, P. J. and TRIVEDI, P. K. (2005). *Microeconometrics*: *Methods and Applications*. Cambridge Univ. Press, Cambridge.

DOWNS, T. D. and MARDIA, K. V. (2002). Circular regression. *Biometrika* **89** 683–697. MR1929172 https://doi.org/10.1093/biomet/89.3.683

FISHER, N. I. and LEE, A. J. (1992). Regression models for an angular response. *Biometrics* **48** 665–677. MR1187598 https://doi.org/10.2307/2532334

HECKMAN, J. (1974). Shadow prices, market wages and labor supply. *Econometrica* **42** 679–694.

HECKMAN, J. (1979). Sample selection bias as a specification error. *J. Roy. Statist. Soc. Ser. B* **49** 127–145.

INGRAHAM, K. A., FEY, N. P., SIMON, A. M. and HARGROVE, L. J. (2016). Assessing the relative contributions of active ankle and knee assistance to the walking mechanics of transfemoral amputees using a powered prosthesis. *PLoS ONE* **11** e0147661. https://doi.org/10.1371/journal.pone.0147661

JAMMALAMADAKA, S. R. and SENGUPTA, A. (2001). *Topics in Circular Statistics*. *Series on Multivariate Analysis* **5**. World Scientific Co., Inc., River Edge, NJ. MR1836122 https://doi.org/10.1142/9789812779267

JAMMALAMADAKA, S. R., BHADRA, N., CHARURVEDI, D., KUTTY, T. K., MUJUMDER, P. P. and PODUVAL, G. (1986). Functional assessment of knee and ankle during level walking. In *Data Analysis in Life Science* (T. Krishnan, ed.) 21–54. Indian Statistical Institute, Calcutta, India.

JESSOP, D. M. and PAIN, M. T. G. (2016). Maximum velocities in flexion and extension actions for sport. *J. Human Kinet.* **50** 37–44.

JHA, J. (2020). Best approach direction for spherical random variables. https://doi.org/10.13140/RG.2.2.34615.42407/1.

Key words and phrases. Abnormal gait, astigmatism, latent variable, Metropolis–Hastings algorithm, rose diagram, truncated wrapped Cauchy.

JHA, J. and BHUYAN, P. (2021). Supplement to "Two-stage circular-circular regression with zero inflation: Application to medical sciences." https://doi.org/10.1214/20-AOAS1429SUPP
JHA, J. and BISWAS, A. (2017). Multiple circular-circular regression. *Stat. Model.* **17** 142–171. MR3648059 https://doi.org/10.1177/1471082X16679501
JHA, J. and BISWAS, A. (2018). Circular-circular regression model with a spike at zero. *Stat. Med.* **37** 71–81. MR3738062 https://doi.org/10.1002/sim.7496
KATO, S., SHIMIZU, K. and SHIEH, G. S. (2008). A circular-circular regression model. *Statist. Sinica* **18** 633–645. MR2432283
LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.
LIPFERT, S. W., GÜNTHER, M., RENJEWSKI, D. and SEYFARTH, A. (2014). Impulsive ankle push-off powers leg swing in human walking. *J. Exp. Biol.* **217** 1218–1228.
MACKENZIE, J. K. (1957). The estimation of an orientation relationship. *Acta Crystallogr.* **10** 61–62. MR0082732
MAHLKNECHT, P., KIECHL, S., BLOEM, B. R., WILLEIT, J., SCHERFLER, C., GASPERI, A., RUNGGER, G., POEWE, W. and SEPPI, K. (2013). Prevalence and burden of gait disorders in elderly men and women aged 60–97 years: A population-based study. *PLoS ONE* **8** e69627. https://doi.org/10.1371/journal.pone.0069627
MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1828667
MIMOUNI, M., NEMET, A., POKROY, R., SELA, T., MUNZER, G. and KAISERMAN, I. (2017). The effect of astigmatism axis on visual acuity. *Eur. J. Ophthalmol.* **27** 308–311. https://doi.org/10.5301/ejo.5000890
MIN, Y. and AGRESTI, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *J. Iran. Stat. Soc.* (*JIRSS*) **1** 7–33.
MOHAN, M. (1989). National survey of blindness—India. NPCB-WHO Report, Ministry of Health and Family Welfare, Government of India, New Delhi.
MORLET, N., MINASSIAN, D. and DART, J. (2001). Astigmatism and the analysis of its surgical correction. *Br. J. Ophthalmol.* **85** 1127–1138.
MURTHY, G. V., GUPTA, S., ELLWEIN, L. B., MUNOZ, S. R., BACHANI, D. and DADA, V. K. (2008a). A population-based eye survey of older adults in a rural district of Rajasthan: I. Central vision impairment, blindness, and cataract surgery. *Ophthalmology* **108** 679–85.
MURTHY, G. V. S., GUPTA, S. K., JOHN, N. and VASHIST, P. (2008b). Current status of cataract blindness and vision 2020: The right to sight initiative in India. *Indian J. Ophthalmol.* **56** 489–494.
NUTT, J. G., HORAK, F. B. and BLOEM, B. R. (2011). Milestones in gait, balance, and falling. *Mov. Disord.* **26** 1166–1174.
RAVINDRAN, P. and GHOSH, S. K. (2011). Bayesian analysis of circular data using wrapped distributions. *J. Stat. Theory Pract.* **5** 547–561. MR2919921 https://doi.org/10.1080/15598608.2011.10483731
RIVEST, L. P. (1997). A decentred predictor for circular-circular regression. *Biometrika* **84** 717–726. MR1603956 https://doi.org/10.1093/biomet/84.3.717
ROAAS, A. and ANDERSSON, G. B. J. (1982). Normal range of motion of the hip, knee and ankle joints in male subjects, 30–40 years of age. *Acta Orthop. Scand.* **53** 205–208.
RUEDA, C., FERNÁNDEZ, M. A., BARRAGÁN, S., MARDIA, K. V. and PEDDADA, S. D. (2016). Circular piecewise regression with applications to cell-cycle data. *Biometrics* **72** 1266–1274. MR3591611 https://doi.org/10.1111/biom.12512
SARMA, Y. R. and JAMMALAMADAKA, S. R. (1993). Circular regression. In *Statistical Sciences and Data Analysis* (*Tokyo*, 1991) 109–128. VSP, Utrecht. MR1337039
SCHWARTZ, J. and GILES, D. E. (2016). Bias-reduced maximum likelihood estimation of the zero-inflated Poisson distribution. *Comm. Statist. Theory Methods* **45** 465–478. MR3447927 https://doi.org/10.1080/03610926.2013.824590
THULASIRAJ, R. D., NIRMALAN, P. K., RAMAKRISHNAN, R., KRISHNADAS, R., MANIMEKALAI, T. K., BABURAJAN, N. P., KATZ, J., TIELSCH, J. M. and ROBIN, A. L. (2003). Blindness and vision impairment in a rural South Indian population: The Aravind comprehensive eye survey. *Ophthalmology* **110** 1491–1498.
TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462 https://doi.org/10.2307/1907382
ZHENG, L., MERRIAM, J. C. and ZAIDER, M. (1997). Astigmatism and visual recovery after 'large incision' extracapsular cataract surgery and 'small' incisions for phakoemulsification. *Trans. Am. Ophthalmol. Soc.* **95** 387–415.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1366–1385
https://doi.org/10.1214/20-AOAS1425

FUNCTION-ON-FUNCTION REGRESSION FOR THE IDENTIFICATION OF EPIGENETIC REGIONS EXHIBITING WINDOWS OF SUSCEPTIBILITY TO ENVIRONMENTAL EXPOSURES

BY MICHELE ZEMPLENYI[1,*], MARK J. MEYER[2], ANDRES CARDENAS[3], MARIE-FRANCE HIVERT[4,‡], SHERYL L. RIFAS-SHIMAN[4,§], HEIKE GIBSON[5,‖], ITAI KLOOG[6], JOEL SCHWARTZ[5,**], EMILY OKEN[4,¶], DAWN L. DEMEO[7], DIANE R. GOLD[5,††] AND BRENT A. COULL[1,†]

[1]*Department of Biostatistics, Harvard T. H. Chan School of Public Health,* [*]*mzemplenyi@g.harvard.edu;* [†]*bcoull@hsph.harvard.edu*

[2]*Department of Mathematics and Statistics, Georgetown University, mjm556@georgetown.edu*

[3]*Division of Environmental Health Sciences, University of California, Berkeley, andres.cardenas@berkeley.edu*

[4]*Department of Population Medicine, Harvard Medical School,* [‡]*mhivert@partners.org;* [§]*sheryl_rifas@harvardpilgrim.org;* [¶]*emily_oken@harvardpilgrim.org*

[5]*Department of Environmental Health, Harvard T. H. Chan School of Public Health,* [‖]*hgibson@hsph.harvard.edu;* [**]*joel@hsph.harvard.edu;* [††]*redrg@hsph.harvard.edu*

[6]*Department of Geography and Environmental Development, Ben-Gurion University, ikloog@gmail.com*

[7]*Center for Chest Diseases, Brigham and Women's Hospital, redld@channing.harvard.edu*

The ability to identify time periods when individuals are most susceptible to exposures as well as the biological mechanisms through which these exposures act is of great public health interest. Growing evidence supports an association between prenatal exposure to air pollution and epigenetic marks, such as DNA methylation, but the timing and gene-specific effects of these epigenetic changes are not well understood. Here, we present the first study that aims to identify prenatal windows of susceptibility to air pollution exposures in cord blood DNA methylation. In particular, we propose a function-on-function regression model that leverages data from nearby DNA methylation probes to identify epigenetic regions that exhibit windows of susceptibility to ambient particulate matter less than 2.5 microns ($PM_{2.5}$). By incorporating the covariance structure among both the multivariate DNA methylation outcome and the time-varying exposure under study, this framework yields greater power to detect windows of susceptibility and greater control of false discoveries than methods that model probes independently. We compare our method to a distributed lag model approach that models DNA methylation in a probe-by-probe manner, both in simulation and by application to motivating data from the Project Viva birth cohort. We identify a window of susceptibility to $PM_{2.5}$ exposure in the middle of the third trimester of pregnancy in an epigenetic region selected based on prior studies of air pollution effects on epigenome-wide methylation.

REFERENCES

BACCARELLI, A., WRIGHT, R. O., BOLLATI, V., TARANTINI, L., LITONJUA, A. A., SUH, H. H., ZANOBETTI, A., SPARROW, D., VOKONAS, P. S. et al. (2009). Rapid DNA methylation changes after exposure to traffic particles. *Am. J. Respir. Crit. Care Med.* **179**.

BOLLATI, V., TARANTINI, L., HU, H., SCHWARTZ, J. D., WRIGHT, R. J., PARK, S. K., SPARROW, D., VOKONAS, P. S., BACCARELLI, A. et al. (2010). Biomarkers of lead exposure and DNA methylation within retrotransposons. *Environ. Health Perspect.* **118**.

BOSE, S., CHIU, Y.-H. M., HSU, H.-H. L., DI, Q., ROSA, M. J., LEE, A., KLOOG, I., WILSON, A., SCHWARTZ, J. et al. (2017). Prenatal nitrate exposure and childhood asthma. Influence of maternal prenatal stress and fetal sex. *Am. J. Respir. Crit. Care Med.* **196**.

Key words and phrases. Functional data analysis, wavelet regression, windows of susceptibility, epigenetics.

BOSE, S., ROSA, M. J., MATHILDA CHIU, Y.-H., LEON HSU, H.-H., DI, Q., LEE, A., KLOOG, I., WILSON, A., SCHWARTZ, J. et al. (2018). Prenatal nitrate air pollution exposure and reduced child lung function: Timing and fetal sex effects. *Environ. Res.* **167** 591–597.

BOSE, S., ROSS, K. R., ROSA, M. J., CHIU, Y.-H. M., JUST, A., KLOOG, I., WILSON, A., THOMPSON, J., SVENSSON, K. et al. (2019). Prenatal particulate air pollution exposure and sleep disruption in preschoolers: Windows of susceptibility. *Environ. Int.* **124**.

BRETON, C., MARSIT, C., FAUSTMAN, E., NADEAU, K., GOODRICH, J., DOLINOY, D., HERBSTMAN, J., HOLLAND, N., LASALLE, J. et al. (2017). Small-magnitude effect sizes in epigenetic end points are important in children's environmental health studies: The children's environmental health and disease prevention research center's epigenetics working group. *Environ. Health Perspect.* **125** 511–526.

CEDERBAUM, J., POUPLIER, M., HOOLE, P. and GREVEN, S. (2016). Functional linear mixed models for irregularly or sparsely sampled data. *Stat. Model.* **16** 67–88. MR3457688 https://doi.org/10.1177/1471082X15617594

CHIU, Y.-H. M., HSU, H.-H. L., COULL, B. A., BELLINGER, D. C., KLOOG, I., SCHWARTZ, J., WRIGHT, R. O. and WRIGHT, R. J. (2016). Prenatal particulate air pollution and neurodevelopment in urban children: Examining sensitive windows and sex-specific associations. *Environ. Int.* **87** 56–65.

CLEMENT, L., DE BEUF, K., THAS, O., VUYLSTEKE, M., IRIZARRY, R. A. and CRAINICEANU, C. M. (2012). Fast wavelet based functional models for transcriptome analysis with tiling arrays. *Stat. Appl. Genet. Mol. Biol.* **11** Art. 4, 38. MR2924207 https://doi.org/10.2202/1544-6115.1726

DADVAND, P., PARKER, J., BELL, M. L., BONZINI, M., BRAUER, M., DARROW, L. A., GEHRING, U., GLINIANAIA, S. V., GOUVEIA, N. et al. (2013). Maternal exposure to particulate air pollution and term birth weight: A multi-country evaluation of effect and heterogeneity. *Environ. Health Perspect.* **121** 267–373. https://doi.org/10.1289/ehp.1205575

DARROW, L. A., KLEIN, M., STRICKLAND, M. J., MULHOLLAND, J. A. and TOLBERT, P. E. (2011). Ambient air pollution and birth weight in full-term infants in Atlanta, 1994–2004. *Environ. Health Perspect.* **119** 731–737. https://doi.org/10.1289/ehp.1002785

DU, P., ZHANG, X., HUANG, C.-C., JAFARI, N., KIBBE, W., HOU, L. and LIN, S. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform.* **11** 587.

FERNÁNDEZ, L., ORDUÃ, L., PÉREZ, M. and ORDUÃ, J. M. (2020). A new approach for the visualization of DNA methylation results. *Comput. Math. Methods* **2** e1043, 6. MR4189300 https://doi.org/10.1002/cmm4.1043

FERRATY, F., VAN KEILEGOM, I. and VIEU, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.* **109** 10–28. MR2922850 https://doi.org/10.1016/j.jmva.2012.02.008

FERRATY, F., LAKSACI, A., TADJ, A. and VIEU, P. (2011). Kernel regression with functional response. *Electron. J. Stat.* **5** 159–171. MR2786486 https://doi.org/10.1214/11-EJS600

FLEISCH, A., RIFAS-SHIMAN, S., KOUTRAKIS, P., SCHWARTZ, J., KLOOG, I., MELLY, S., COULL, B., ZANOBETTI, A., GILLMAN, M. et al. (2015). Prenatal exposure to traffic pollution: Associations with reduced fetal growth and rapid infant weight gain. *Epidemiology* **26** 43–50.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) 169–193. Oxford Univ. Press, New York. MR1380276

GOLDSMITH, J., BOBB, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011). Penalized functional regression. *J. Comput. Graph. Statist.* **20** 830–851. MR2878950 https://doi.org/10.1198/jcgs.2010.10007

GREVEN, S. and SCHEIPL, F. (2017). A general framework for functional regression modelling. *Stat. Model.* **17** 1–35. MR3619335 https://doi.org/10.1177/1471082X16681317

GRUZIEVA, O., KOGEVINAS, M., RUIZ, J. L., BUSTAMANTE PINEDA, M., ANTÓ I BOQUÉ, J. M., SUNYER DEU, J., VRIJHEID, M., HERNANDEZ FERRE, C. and MELÉN, E. (2019). Prenatal particulate air pollution and DNA methylation in newborns: An epigenome-wide meta-analysis. *Environ. Health Perspect.* **127**.

GUO, S., DIEP, D., PLONGTHONGKUM, N., FUNG, H.-L., ZHANG, K. and ZHANG, K. (2017). Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49** 635–642.

HANCOCK, D. B., EIJGELSHEIM, M., WILK, J. B., GHARIB, S. A., LOEHR, L. R., MARCIANTE, K. D., FRANCESCHINI, N., DURME, Y. M. T. A. V., CHEN, T.-H. et al. (2009). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42** 45.

HARRIS, M. H., GOLD, D. R., RIFAS-SHIMAN, S. L., MELLY, S. J., ZANOBETTI, A., COULL, B. A., SCHWARTZ, J. D., GRYPARIS, A., KLOOG, I. et al. (2016). Prenatal and childhood traffic-related air pollution exposure and childhood executive function and behavior. *Neurotoxicol. Teratol.* **57** 60–70. https://doi.org/10.1016/j.ntt.2016.06.008

HILL, M. (2019). Embryology fetal development. Available at https://embryology.med.unsw.edu.au/embryology/index.php/Fetal_Development, Last accessed on 2019-10-11.

HOBBS, B. D., JONG, K. D., LAMONTAGNE, M., BOSSÉ, Y., SHRINE, N., ARTIGAS, M. S., WAIN, L. V., HALL, I. P., JACKSON, V. E. et al. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* **49**.

HSU, L., SELF, S., GROVE, D., RANDOLPH, T., WANG, K., DELROW, J., LOO, L. and PORTER, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6** 211–226.

HSU, H.-H. L., CHIU, Y.-H. M., COULL, B. A., KLOOG, I., SCHWARTZ, J., LEE, A., WRIGHT, R. O. and WRIGHT, R. J. (2015). Prenatal particulate air pollution and asthma onset in urban children. Identifying sensitive windows and sex differences. *Am. J. Respir. Crit. Care Med.* **192**.

IVANESCU, A. E. (2018). Function-on-function regression for two-dimensional functional data. *Comm. Statist. Simulation Comput.* **47** 2656–2669. MR3863111 https://doi.org/10.1080/03610918.2017.1353619

JOHNSTONE, I. M. and SILVERMAN, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59** 319–351. MR1440585 https://doi.org/10.1111/1467-9868.00071

KLOOG, I., KOUTRAKIS, P., COULL, B., LEE, H. and SCHWARTZ, J. (2011). Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **45** 6267–6275.

KLOOG, I., CHUDNOVSKY, A. A., JUST, A. C., NORDIO, F., KOUTRAKIS, P., COULL, B. A., LYAPUSTIN, A., WANG, Y. and SCHWARTZ, J. (2014). A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmos. Environ.* **95** 581–590.

LAMICHHANE, D. K., RYU, J., LEEM, J.-H., HA, M., HONG, Y.-C., PARK, H., KIM, Y., JUNG, D.-Y., LEE, J.-Y. et al. (2018). Air pollution exposure during pregnancy and ultrasound and birth measures of fetal growth: A prospective cohort study in Korea. *Sci. Total Environ.* **619–620** 834–841.

LAVIGNE, E., DONELLE, J., HATZOPOULOU, M., VAN RYSWYK, K., VAN DONKELAAR, A., MARTIN, R. V., CHEN, H., STIEB, D. M., GASPARRINI, A. et al. (2019). Spatiotemporal variations in ambient ultrafine particles and the incidence of childhood asthma. *Am. J. Respir. Crit. Care Med.* **199**.

LEE, W. and MORRIS, J. S. (2016). Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics* **32** 664–672.

LEE, K. H., TADESSE, M. G., BACCARELLI, A. A., SCHWARTZ, J. and COULL, B. A. (2017). Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics* **73** 232–241. MR3632369 https://doi.org/10.1111/biom.12557

LEE, A., LEGRAND, B., HSU, H., CHIU, Y., BRENNAN, K., BOSE, S., ROSA, M., KLOOG, I., WILSON, A. et al. (2018). Prenatal fine particulate exposure associated with reduced childhood lung function and nasal epithelia GSTP1 hypermethylation: Sex-specific effects. *Am. J. Respir. Crit. Care Med.* **197**.

LEE, W., MIRANDA, M. F., RAUSCH, P., BALADANDAYUTHAPANI, V., FAZIO, M., DOWNS, J. C. and MORRIS, J. S. (2019). Bayesian semiparametric functional mixed models for serially correlated functional data with application to Glaucoma data. *J. Amer. Statist. Assoc.* **114** 495–513. MR3963158 https://doi.org/10.1080/01621459.2018.1476242

LEPEULE, J., BACCARELLI, A., TARANTINI, L., MOTTA, V., CANTONE, L., LITONJUA, A. A., SPARROW, D., VOKONAS, P. S. and SCHWARTZ, J. (2012). Gene promoter methylation is associated with lung function in the elderly: The normative aging study. *Epigenetics* **7** 261–269.

LI, X., HAWKINS, G. A., AMPLEFORD, E. J., MOORE, W. C., LI, H., HASTIE, A. T., HOWARD, T. D., BOUSHEY, H. A., BUSSE, W. W. et al. (2013). Genome-wide association study identifies TH1 pathway genes associated with lung function in asthmatic patients. *The Journal of Allergy and Clinical Immunology* **132** 313–320.

MALFAIT, N. and RAMSAY, J. O. (2003). The historical functional linear model. *Canad. J. Statist.* **31** 115–128. MR2016223 https://doi.org/10.2307/3316063

MALLOY, E. J., MORRIS, J. S., ADAR, S. D., SUH, H., GOLD, D. R. and COULL, B. A. (2010). Wavelet-based functional linear mixed models: An application to measurement error–corrected distributed lag models. *Biostatistics* **11** 432–452.

MEYER, M. J., COULL, B. A., VERSACE, F., CINCIRIPINI, P. and MORRIS, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics* **71** 563–574. MR3402592 https://doi.org/10.1111/biom.12299

MITRA, A. and SONG, J. (2012). WaveSeq: A novel data-driven method of detecting histone modification enrichments using wavelets (ChIP-seq and wavelets) **7**.

MORRIS, J. S. (2015). Functional regression. *Annual Reviews of Statistics and Its Application* **2** 321–359.

MORRIS, J. S. (2017). Comparison and contrast of two general functional regression modelling frameworks [Discussion of MR3619335]. *Stat. Model.* **17** 59–85. MR3619339 https://doi.org/10.1177/1471082X16681875

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981 https://doi.org/10.1111/j.1467-9868.2006.00539.x

MORRIS, J. S., BROWN, P. J., HERRICK, R. C., BAGGERLY, K. A. and COOMBES, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489, 667. MR2432418 https://doi.org/10.1111/j.1541-0420.2007.00895.x

MÜLLER, P., PARMIGIANI, G., RICE, V. C., FERNÁNDEZ-VAL, I. and KOWALSKI, A. (2006). FDR and Bayesian multiple comparison rules. Working paper.

NGUYEN, N., VO, A. and WON, K. (2014). A wavelet-based method to exploit epigenomic language in the regulatory region. *Bioinformatics* **30** 908–914.

OKEN, E., BACCARELLI, A. A., GOLD, D. R., KLEINMAN, K. P., LITONJUA, A. A., MEO, D. D., RICH-EDWARDS, J. W., RIFAS-SHIMAN, S. L., SAGIV, S. et al. (2015). Cohort profile: Project viva. *Int. J. Epidemiol.* **44** 37–48. https://doi.org/10.1093/ije/dyu008

RAMSAY, J. O. and DALZELL, C. J. (1991). Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B* **53** 539–572. MR1125714

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

REISS, P. T., GOLDSMITH, J., SHANG, H. L. and OGDEN, R. T. (2017). Methods for scalar-on-function regression. *Int. Stat. Rev.* **85** 228–249. MR3686566 https://doi.org/10.1111/insr.12163

SARDY, S., PERCIVAL, D., BRUCE, A., GAO, H.-Y. and STUETZLE, W. (1999). Wavelet shrinkage for unequally spaced data. *Stat. Comput.* **9** 65–75.

SCHEIPL, F., GERTHEISS, J. and GREVEN, S. (2016). Generalized functional additive mixed models. *Electron. J. Stat.* **10** 1455–1492. MR3507370 https://doi.org/10.1214/16-EJS1145

SCHEIPL, F., STAICU, A.-M. and GREVEN, S. (2015). Functional additive mixed models. *J. Comput. Graph. Statist.* **24** 477–501. MR3357391 https://doi.org/10.1080/10618600.2014.901914

SCHNEIDER, J. S., KIDD, S. K. and ANDERSON, D. W. (2013). Influence of developmental lead exposure on expression of DNA methyltransferases and methyl cytosine-binding proteins in hippocampus. *Toxicol Lett* **217** 75–81. https://doi.org/10.1016/j.toxlet.2012.12.004

SCHWARTZ, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology* **11** 320–326.

SHAH, P. S. and BALKHAIR, T. (2011). Air pollution and birth outcomes: A systematic review. *Environ. Int.* **37** 498–516. https://doi.org/10.1016/j.envint.2010.10.009

SOBERANES, S., GONZALEZ, A., URICH, D., CHIARELLA, S. E., RADIGAN, K. A., OSORNIO-VARGAS, A., JOSEPH, J., KALYANARAMAN, B., RIDGE, K. M. et al. (2012). Particulate matter air pollution induces hypermethylation of the p16 promoter via a mitochondrial ROS-JNK-DNMT1 pathway. *Sci. Rep.* **2**.

SORDILLO, J. E., RIFAS-SHIMAN, S. L., SWITKOWSKI, K., COULL, B., GIBSON, H., RICE, M., PLATTS-MILLS, T. A. E., KLOOG, I., LITONJUA, A. A. et al. (2019). Prenatal oxidative balance and risk of asthma and allergic disease in adolescence. *The Journal of Allergy and Clinical Immunology*.

VAN ROSSEM, L., RIFAS-SHIMAN, S. L., MELLY, S. J., KLOOG, I., LUTTMANN-GIBSON, H., ZANOBETTI, A., COULL, B. A., SCHWARTZ, J. D., MITTLEMAN, M. A. et al. (2015). Prenatal air pollution exposure and newborn blood pressure. *Environ. Health Perspect.* **123** 353–359. https://doi.org/10.1289/ehp.1307419

WAND, M. P. and ORMEROD, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electron. J. Stat.* **5** 1654–1717. MR2870147 https://doi.org/10.1214/11-EJS652

WANG, W. (2014). Linear mixed function-on-function regression models. *Biometrics* **70** 794–801. MR3295740 https://doi.org/10.1111/biom.12207

WARREN, J. L., KONG, W., LUBEN, T. J. and CHANG, H. H. (2020). Critical window variable selection: Estimating the impact of air pollution on very preterm birth. *Biostatistics* **21** 790–806. MR4164058 https://doi.org/10.1093/biostatistics/kxz006

WILSON, A., CHIU, Y.-H. M., HSU, H.-H. L., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2017). Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health. *Biostatistics* **18** 537–552. MR3799593 https://doi.org/10.1093/biostatistics/kxx002

WU, H., JIANG, B., ZHU, P., GENG, X., LIU, Z., CUI, L. and YANG, L. (2018). Associations between maternal weekly air pollutant exposures and low birth weight: A distributed lag non-linear model. *Environ. Res. Lett.* **13**.

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660

ZANOBETTI, A., WAND, M. P., SCHWARTZ, J. and RYAN, L. M. (2000). Generalized additive distributed lag models: Quantifying mortality displacement. *Biostatistics* **1** 279–292.

ZEMPLENYI, M., MEYER, M. J., CARDENAS, A., HIVERT, M. F., RIFAS-SHIMAN, S. L., GIBSON, H., KLOOG, I., SCHWARTZ, J., OKEN, E. et al. (2021). Supplement to "Function-on-function regression for the identification of epigenetic regions exhibiting windows of susceptibility to environmental exposures." https://doi.org/10.1214/20-AOAS1425SUPPA, https://doi.org/10.1214/20-AOAS1425SUPPB.

ZHANG, Y., SHIN, H., SONG, J. S., LEI, Y. and LIU, X. S. (2008). Identifying positioned nucleosomes with epigenetic marks in human from ChIP-seq. *BMC Genomics* **9** 1–11.

ZHU, H., VERSACE, F., CINCIRIPINI, P. and MORRIS, J. S. (2018). Robust functional mixed models for spatially correlated functional regression, with application to event-related potentials for nicotine-addicted individuals. *NeuroImage* **181** 501–512.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1386–1404
https://doi.org/10.1214/20-AOAS1435

PERTURBED FACTOR ANALYSIS: ACCOUNTING FOR GROUP DIFFERENCES IN EXPOSURE PROFILES

BY ARKAPRAVA ROY[1], ISAAC LAVINE[2,*], AMY H. HERRING[2,†] AND DAVID B. DUNSON[2,‡]

[1]*Department of Biostatistics, University of Florida, ark007@ufl.edu*
[2]*Department of Statistical Science, Duke University, *isaac.lavine@duke.edu; †amy.herring@duke.edu; ‡dunson@duke.edu*

In this article we investigate group differences in phthalate exposure profiles using NHANES data. Phthalates are a family of industrial chemicals used in plastics and as solvents. There is increasing evidence of adverse health effects of exposure to phthalates on reproduction and neurodevelopment and concern about racial disparities in exposure. We would like to identify a single set of low-dimensional factors summarizing exposure to different chemicals, while allowing differences across groups. Improving on current multigroup additive factor models, we propose a class of Perturbed Factor Analysis (PFA) models that assume a common factor structure after perturbing the data via multiplication by a group-specific matrix. Bayesian inference algorithms are defined using a matrix normal hierarchical model for the perturbation matrices. The resulting model is just as flexible as current approaches in allowing arbitrarily large differences across groups but has substantial advantages that we illustrate in simulation studies. Applying PFA to NHANES data, we learn common factors summarizing exposures to phthalates, while showing clear differences across groups.

REFERENCES

ASSMANN, C., BOYSEN-HOGREFE, J. and PAPE, M. (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *J. Econometrics* **192** 190–206. MR3463672 https://doi.org/10.1016/j.jeconom.2015.10.010

BENJAMIN, S., MASAI, E., KAMIMURA, N., TAKAHASHI, K., ANDERSON, R. C. and FAISAL, P. A. (2017). Phthalates impact human health: Epidemiological evidences and plausible mechanism of action. *J. Hazard. Mater.* **340** 360–383. https://doi.org/10.1016/j.jhazmat.2017.06.036

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

BLOOM, M. S., WENZEL, A. G., BROCK, J. W., KUCKLICK, J. R., WINELAND, R. J., CRUZE, L., UNAL, E. R., YUCEL, R. M., JIYESSOVA, A. et al. (2019). Racial disparity in maternal phthalates exposure; association with racial disparity in fetal growth and birth outcomes. *Environ. Int.* **127** 473–486.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869

DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2018). Bayesian multi-study factor analysis for high-throughput biological data. ArXiv preprint. Available at arXiv:1806.09896.

DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. MR3953734 https://doi.org/10.1111/biom.12974

DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. MR3584158 https://doi.org/10.1016/j.spl.2016.11.014

FENG, Q., HANNIG, J. and MARRON, J. S. (2015). Non-iterative joint and individual variation explained. ArXiv preprint. Available at arXiv:1512.04060.

FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008

Key words and phrases. Bayesian, chemical mixtures, factor analysis, hierarchical model, metaanalysis, perturbation matrix, phthalate exposures, racial disparities.

FRÜEHWIRTH-SCHNATTER, S. and LOPES, H. F. (2018). Sparse Bayesian factor analysis when the number of factors is unknown. ArXiv preprint. Available at arXiv:1804.04231.

JAMES-TODD, T. M., MEEKER, J. D., HUANG, T., HAUSER, R., SEELY, E. W., FERGUSON, K. K., RICH-EDWARDS, J. W. and MCELRATH, T. F. (2017). Racial and ethnic variations in phthalate metabolite concentration changes across full-term pregnancies. *J. Expo. Sci. Environ. Epidemiol.* **27** 160–166. https://doi.org/10.1038/jes.2016.2

KIM, S. H. and PARK, M. J. (2014). Phthalate exposure and childhood obesity. *Ann. Pediatr. Endocrinol. Metab.* **19** 69.

KIM, S., KANG, D., HUO, Z., PARK, Y. and TSENG, G. C. (2018). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* **34** 1321–1328.

KINGMA, D. P. and WELLING, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

LAWRENCE, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems* 16 (S. Thrun, L. K. Saul and B. Schölkopf, eds.) 329–336. MIT Press, Cambridge.

LAWRENCE, N. and CANDELA, J. Q. (2006). Local distance preservation in the GP-LVM through back constraints. In *International Conference on Machine Learning*'06.

LEE, S. X., LIN, T.-I. and MCLACHLAN, G. J. (2018). Mixtures of factor analyzers with fundamental skew symmetric distributions. ArXiv preprint. Available at arXiv:1802.02467.

LI, G. and JUNG, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics* **73** 1433–1442. MR3744555 https://doi.org/10.1111/biom.12698

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 https://doi.org/10.1214/12-AOAS597

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

MARESCA, M. M., HOEPNER, L. A., HASSOUN, A., OBERFIELD, S. E., MOONEY, S. J., CALAFAT, A. M., RAMIREZ, J., FREYER, G., PERERA, F. P. et al. (2016). Prenatal exposure to phthalates and childhood body size in an urban cohort. *Environ. Health Perspect.* **124** 514–520.

MCPARLAND, D., GORMLEY, I. C., MCCORMICK, T. H., CLARK, S. J., KABUDULA, C. W. and COLLINSON, M. A. (2014). Clustering South African households based on their asset status using latent variable models. *Ann. Appl. Stat.* **8** 747–776. MR3262533 https://doi.org/10.1214/14-AOAS726

MURPHY, K., VIROLI, C. and GORMLEY, I. C. (2020a). Infinite mixtures of infinite factor analysers. *Bayesian Anal.* **15** 937–963. MR4132655 https://doi.org/10.1214/19-BA1179

MURPHY, K., VIROLI, C. and GORMLEY, I. C. (2020b). `IMIFA`: Infinite mixtures of infinite factor analysers and related models. R package version 2.1.3.

NEUWIRTH, E. (2014). RColorBrewer: ColorBrewer palettes. R package version 1.1-2.

PU, Y., GAN, Z., HENAO, R., YUAN, X., LI, C., STEVENS, A. and CARIN, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems* 2352–2360.

ROČKOVÁ, V. and GEORGE, E. I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *J. Amer. Statist. Assoc.* **111** 1608–1622. MR3601721 https://doi.org/10.1080/01621459.2015.1100620

ROY, A., SCHAICH-BORG, J. and DUNSON, D. B. (2019). Bayesian time-aligned factor analysis of paired multivariate time series. ArXiv preprint. Available at arXiv:1904.12103.

ROY, A., LAVINE, I., HERRING, A. H. and DUNSON, D. B. (2021). Supplement to "Perturbed factor analysis: Accounting for group differences in exposure profiles." https://doi.org/10.1214/20-AOAS1435SUPPA, https://doi.org/10.1214/20-AOAS1435SUPPB

SARKAR, A., PATI, D., CHAKRABORTY, A., MALLICK, B. K. and CARROLL, R. J. (2018). Bayesian semiparametric multivariate density deconvolution. *J. Amer. Statist. Assoc.* **113** 401–416. MR3803474 https://doi.org/10.1080/01621459.2016.1260467

SEBER, G. A. (2009). *Multivariate Observations* **252**. Wiley, New York.

TAYLOR, K. W., TROESTER, M. A., HERRING, A. H., ENGEL, L. S., NICHOLS, H. B., SANDLER, D. P. and BAIRD, D. D. (2018). Associations between personal care product use patterns and breast cancer risk among white and black women in the sister study. *Environ. Health Perspect.* **126** 027011. https://doi.org/10.1289/EHP1480

WEISSENBURGER-MOSER, L., MEZA, J., YU, F., SHIYANBOLA, O., ROMBERGER, D. J. and LEVAN, T. D. (2017). A principal factor analysis to characterize agricultural exposures among Nebraska veterans. *J. Expo. Sci. Environ. Epidemiol.* **27** 214–220. https://doi.org/10.1038/jes.2016.20

ZHANG, Y., MENG, X., CHEN, L., LI, D., ZHAO, L., ZHAO, Y., LI, L. and SHI, H. (2014). Age and sex-specific relationships between phthalate exposures and obesity in Chinese children at puberty. *PLoS ONE* **9** e104852.

ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc*. **110** 1562–1576. MR3449055 https://doi.org/10.1080/01621459.2014.983233

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1405–1430
https://doi.org/10.1214/21-AOAS1461

BAYESIAN JOINT MODELING OF CHEMICAL STRUCTURE AND DOSE RESPONSE CURVES

BY KELLY R. MORAN[1,*], DAVID DUNSON[1,†], MATTHEW W. WHEELER[2] AND AMY H. HERRING[1,‡]

[1]*Department of Statistical Science, Duke University,* [*]*kelly.r.moran@duke.edu;* [†]*dunson@duke.edu;* [‡]*amy.herring@duke.edu*
[2]*Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, matt.wheeler@nih.gov*

Today there are approximately 85,000 chemicals regulated under the Toxic Substances Control Act, with around 2,000 new chemicals introduced each year. It is impossible to screen all of these chemicals for potential toxic effects, either via full organism *in vivo* studies or *in vitro* high-throughput screening (HTS) programs. Toxicologists face the challenge of choosing which chemicals to screen, and predicting the toxicity of as yet unscreened chemicals. Our goal is to describe how variation in chemical structure relates to variation in toxicological response to enable *in silico* toxicity characterization designed to meet both of these challenges. With our Bayesian partially Supervised Sparse and Smooth Factor Analysis (BS^{3}FA) model, we learn a distance between chemicals targeted to toxicity, rather than one based on molecular structure alone. Our model also enables the prediction of chemical dose-response profiles based on chemical structure (i.e., without *in vivo* or *in vitro* testing) by taking advantage of a large database of chemicals that have already been tested for toxicity in HTS programs. We show superior simulation performance in distance learning and modest to large gains in predictive ability compared to existing methods. Results from the high-throughput screening data application elucidate the relationship between chemical structure and a toxicity-relevant high-throughput assay. An R package for BS^{3}FA is available online at https://github.com/kelrenmor/bs3fa.

REFERENCES

BARBER, R. F., REIMHERR, M. and SCHILL, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electron. J. Stat.* **11** 1351–1389. MR3635916 https://doi.org/10.1214/17-EJS1260

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

CANALE, A. and DUNSON, D. B. (2013). Nonparametric Bayes modelling of count processes. *Biometrika* **100** 801–816. MR3142333 https://doi.org/10.1093/biomet/ast037

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

CHEN, Y., GOLDSMITH, J. and OGDEN, R. T. (2016). Variable selection in function-on-scalar regression. *Stat* **5** 88–101. MR3478799 https://doi.org/10.1002/sta4.106

DHALIWAL, L. K., SURI, V., GUPTA, K. R. and SAHDEV, S. (2011). Tamoxifen: An alternative to clomiphene in women with polycystic ovary syndrome. *Journal of Human Reproductive Sciences* **4** 76.

DIX, D. J., HOUCK, K. A., MARTIN, M. T., RICHARD, A. M., SETZER, R. W. and KAVLOCK, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **95** 5–12. https://doi.org/10.1093/toxsci/kfl103

DOCAMPO, R. and MORENO, S. N. (1990). The metabolism and mode of action of gentian violet. *Drug Metab. Rev.* **22** 161–178.

DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. MR3584158 https://doi.org/10.1016/j.spl.2016.11.014

FAN, Z. and REIMHERR, M. (2017). High-dimensional adaptive function-on-scalar regression. *Econom. Stat.* **1** 167–183. MR3669995 https://doi.org/10.1016/j.ecosta.2016.08.001

Key words and phrases. Dimension reduction, distance learning, functional prediction, high-throughput screening, toxicity, ToxCast, QSAR.

HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J. Amer. Statist. Assoc.* **110** 435–448. MR3338514 https://doi.org/10.1080/01621459.2014.993077

HONG, H., XIE, Q., GE, W., QIAN, F., FANG, H., SHI, L., SU, Z., PERKINS, R. and TONG, W. (2008). Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **48** 1337–1344.

HONG, H., SLAVOV, S., GE, W., QIAN, F., SU, Z., FANG, H., CHENG, Y., PERKINS, R., SHI, L. et al. (2012). Mold2 molecular descriptors for QSAR. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* **2** 65–109.

JUDSON, R. S., HOUCK, K. A., KAVLOCK, R. J., KNUDSEN, T. B., MARTIN, M. T., MORTENSEN, H. M., REIF, D. M., ROTROFF, D. M., SHAH, I. et al. (2009). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environ. Health Perspect.* **118** 485–492.

KAVLOCK, R., CHANDLER, K., HOUCK, K., HUNTER, S., JUDSON, R., KLEINSTREUER, N., KNUDSEN, T., MARTIN, M., PADILLA, S. et al. (2012). Update on EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* **25** 1287–1302.

KLIEWER, S. A., GOODWIN, B. and WILLSON, T. M. (2002). The nuclear pregnane X receptor: A key regulator of xenobiotic metabolism. *Endocr. Rev.* **23** 687–702. https://doi.org/10.1210/er.2001-0038

KNOWLES, D. and GHAHRAMANI, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann. Appl. Stat.* **5** 1534–1552. MR2849785 https://doi.org/10.1214/10-AOAS435

KOWAL, D. R. and BOURGEOIS, D. C. (2020). Bayesian function-on-scalars regression for high-dimensional data. *J. Comput. Graph. Statist.* **29** 629–638. MR4153187 https://doi.org/10.1080/10618600.2019.1710837

LI, G., SHEN, H. and HUANG, J. Z. (2016). Supervised sparse and functional principal component analysis. *J. Comput. Graph. Statist.* **25** 859–878. MR3533642 https://doi.org/10.1080/10618600.2015.1064434

LIU, R., RALLO, R., GEORGE, S., JI, Z., NAIR, S., NEL, A. E. and COHEN, Y. (2011). Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small* **7** 1118–1126.

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 https://doi.org/10.1214/12-AOAS597

LOW-KAM, C., TELESCA, D., JI, Z., ZHANG, H., XIA, T., ZINK, J. I. and NEL, A. E. (2015). A Bayesian regression tree approach to identify the effect of nanoparticles' properties on toxicity profiles. *Ann. Appl. Stat.* **9** 383–401. MR3341120 https://doi.org/10.1214/14-AOAS797

MAKALIC, E. and SCHMIDT, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* **23** 179–182.

MARTIN, Y. C., KOFRON, J. L. and TRAPHAGEN, L. M. (2002). Do structurally similar molecules have similar biological activity? *J. of Med. Chem.* **45** 4350–4358.

MENG, J., ZHANG, J., QI, Y., CHEN, Y. and HUANG, Y. (2010). Uncovering transcriptional regulatory networks by sparse Bayesian factor model. *EURASIP J. Adv. Signal Process.* **2010** 3.

MEYER, M. J., COULL, B. A., VERSACE, F., CINCIRIPINI, P. and MORRIS, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics* **71** 563–574. MR3402592 https://doi.org/10.1111/biom.12299

MONTAGNA, S., TOKDAR, S. T., NEELON, B. and DUNSON, D. B. (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics* **68** 1064–1073. MR3040013 https://doi.org/10.1111/j.1541-0420.2012.01788.x

MORAN, K. R., DUNSON, D., WHEELER, M. W. and HERRING, A. H. (2021). Supplement to "Bayesian joint modeling of chemical structure and dose response curves." https://doi.org/10.1214/21-AOAS1461SUPPA, https://doi.org/10.1214/21-AOAS1461SUPPB, https://doi.org/10.1214/21-AOAS1461SUPPC

NIKOLOVA, N. and JAWORSKA, J. (2003). Approaches to measure chemical similarity—a review. *QSAR & Combinatorial Science* **22** 1006–1026.

O'CONNELL, M. J. and LOCK, E. F. (2019). Linked matrix factorization. *Biometrics* **75** 582–592. MR3999181 https://doi.org/10.1111/biom.13010

PATEL, T., TELESCA, D., LOW-KAM, C., JI, Z. X., ZHANG, H. Y., XIA, T., ZINC, J. I. and NEL, A. E. (2014). Relating nano-particle properties to biological outcomes in exposure escalation experiments. *Environmetrics* **25** 57–68. MR3233744 https://doi.org/10.1002/env.2246

PATI, D., BHATTACHARYA, A., PILLAI, N. S. and DUNSON, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. MR3210997 https://doi.org/10.1214/14-AOS1215

RAY, P., ZHENG, L., LUCAS, J. and CARIN, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* **30** 1370–1376.

SEYEDOSHOHADAEI, F., ZANDVAKILY, F. and SHAHGEIBI, S. (2012). Comparison of the effectiveness of clomiphene citrate, tamoxifen and letrozole in ovulation induction in infertility due to isolated unovulation. *Iran. J. Reprod. Med.* **10** 531–536.

SRIVASTAVA, S., SINHA, R. and ROY, D. (2004). Toxicological effects of malachite green. *Aquat. Toxicol.* **66** 319–329.

WEININGER, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** 31–36.

WHEELER, M. W. (2019). Bayesian additive adaptive basis tensor product models for modeling high dimensional surfaces: An application to high-throughput toxicity testing. *Biometrics* **75** 193–201. MR3953720

WILSON, A., REIF, D. M. and REICH, B. J. (2014). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics* **70** 237–246. MR3251684 https://doi.org/10.1111/biom.12114

YOSHIDA, R. and WEST, M. (2010). Bayesian learning in sparse graphical factor models via variational mean-field annealing. *J. Mach. Learn. Res.* **11** 1771–1798. MR2653356

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1431–1454
https://doi.org/10.1214/21-AOAS1466

SIMULTANEOUS NON-GAUSSIAN COMPONENT ANALYSIS (SING) FOR DATA INTEGRATION IN NEUROIMAGING

BY BENJAMIN B. RISK[1] AND IRINA GAYNANOVA[2]

[1]*Department of Biostatistics and Bioinformatics, Emory University, benjamin.risk@emory.edu*
[2]*Department of Statistics, Texas A&M University, irinag@stat.tamu.edu*

As advances in technology allow the acquisition of complementary information, it is increasingly common for scientific studies to collect multiple datasets. Large-scale neuroimaging studies often include multiple modalities (e.g., task functional MRI, resting-state fMRI, diffusion MRI, and/or structural MRI) with the aim to understand the relationships between datasets. In this study, we seek to understand whether regions of the brain activated in a working memory task relate to resting-state correlations. In neuroimaging, a popular approach uses principal component analysis for dimension reduction prior to canonical correlation analysis with joint independent component analysis, but this may discard biological features with low variance and/or spuriously associate structure unique to a dataset with joint structure. We introduce SImultaneous Non-Gaussian component analysis (SING) in which dimension reduction and feature extraction are achieved simultaneously, and shared information is captured via subject scores. We apply our method to a working memory task and resting-state correlations from the Human Connectome Project. We find joint structure as evident from joint scores whose loadings highlight resting-state correlations involving regions associated with working memory. Moreover, some of the subject scores are related to fluid intelligence.

REFERENCES

AKIKI, T. J. and ABDALLAH, C. G. (2019). Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Sci. Rep.* **9** 1–15.

AMICO, E., MARINAZZO, D., DI PERRI, C., HEINE, L., ANNEN, J., MARTIAL, C., DZEMIDZIC, M., KIRSCH, M., BONHOMME, V. et al. (2017). Mapping the functional connectome traits of levels of consciousness. *NeuroImage* **148** 201–211.

ARGELAGUET, R., VELTEN, B., ARNOL, D., DIETRICH, S., ZENZ, T., MARIONI, J. C., BUETTNER, F., HUBER, W. and STEGLE, O. (2018). Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14** e8124. https://doi.org/10.15252/msb.20178124

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

BECKMANN, C. F. and SMITH, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imag.* **23** 137–152.

BELL, A. J. and SEJNOWSKI, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7** 1129–1159.

BICKEL, P. J., KUR, G. and NADLER, B. (2018). Projection pursuit in high dimensions. *Proc. Natl. Acad. Sci. USA* **115** 9151–9156. MR3856112 https://doi.org/10.1073/pnas.1801177115

BLANCHARD, G., SUGIYAMA, M., KAWANABE, M., SPOKOINY, V. and MÜLLER, K.-R. (2005). Non-Gaussian component analysis: A semi-parametric framework for linear dimension reduction. In *Advances in Neural Information Processing Systems* 131–138.

BRESSLER, S. L. and MENON, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends Cogn. Sci.* **14** 277–290.

Key words and phrases. Canonical correlation analysis, data fusion, independent component analysis, JIVE, multiblock, multimodality, multiview, projection pursuit, unsupervised learning.

CALHOUN, V. D. and ADALI, T. (2012). Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* **5** 60–73. https://doi.org/10.1109/RBME.2012.2211076

CALHOUN, V. D., LIU, J. and ADALI, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45** S163–S172.

CALHOUN, V. D. and SUI, J. (2016). Multimodal fusion of brain imaging data: A key to finding the missing link (s) in complex mental illness. *Biological Psychiatry* **1** 230–244.

CALHOUN, V. D., ADALI, T., GIULIANI, N. R., PEKAR, J. J., KIEHL, K. A. and PEARLSON, G. D. (2006a). Method for multimodal analysis of independent source differences in schizophrenia: Combining gray matter structural and auditory oddball functional data. *Hum. Brain Mapp.* **27** 47–62. https://doi.org/10.1002/hbm.20166

CALHOUN, V. D., ADALI, T., KIEHL, K. A., ASTUR, R., PEKAR, J. J. and PEARLSON, G. D. (2006b). A method for multitask fMRI data fusion applied to schizophrenia. *Hum. Brain Mapp.* **27** 598–610. https://doi.org/10.1002/hbm.20204

CRESPI, C., GALANDRA, C., MANERA, M., BASSO, G., POGGI, P. and CANESSA, N. (2019). Executive impairment in alcohol use disorder reflects structural changes in large-scale brain networks: A joint independent component analysis on gray-matter and white-matter features. *Front. Psychol.* **10** 2479. https://doi.org/10.3389/fpsyg.2019.02479

FENG, Q., JIANG, M., HANNIG, J. and MARRON, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.* **166** 241–265. MR3799646 https://doi.org/10.1016/j.jmva.2018.03.008

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **100** 881–890.

GAYNANOVA, I. and LI, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75** 1121–1132. MR4041816 https://doi.org/10.1111/biom.13108

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80** 105–124.

GLASSER, M. F., SMITH, S. M., MARCUS, D. S., ANDERSSON, J. L., AUERBACH, E. J., BEHRENS, T. E., COALSON, T. S., HARMS, M. P., JENKINSON, M. et al. (2016a). The human connectome project's neuroimaging approach. *Nat. Neurosci.* **19** 1175–1187.

GLASSER, M. F., COALSON, T. S., ROBINSON, E. C., HACKER, C. D., HARWELL, J., YACOUB, E., UGURBIL, K., ANDERSSON, J., BECKMANN, C. F. et al. (2016b). A multi-modal parcellation of human cerebral cortex. *Nature* **536** 171–178.

GOULDEN, N., KHUSNULINA, A., DAVIS, N. J., BRACEWELL, R. M., BOKDE, A. L., MCNULTY, J. P. and MULLINS, P. G. (2014). The salience network is responsible for switching between the default mode network and the central executive network: Replication from DCM. *NeuroImage* **99** 180–190.

GROVES, A. R., BECKMANN, C. F., SMITH, S. M. and WOOLRICH, M. W. (2011). Linked independent component analysis for multimodal data fusion. *NeuroImage* **54** 2198–2217.

HINAULT, T., LARCHER, K., ZAZUBOVITS, N., GOTMAN, J. and DAGHER, A. (2019). Spatio-temporal patterns of cognitive control revealed with simultaneous electroencephalography and functional magnetic resonance imaging. *Hum. Brain Mapp.* **40** 80–97. https://doi.org/10.1002/hbm.24356

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

HYVÄRINEN, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10** 626–634.

JAEGGI, S. M., BUSCHKUEHL, M., PERRIG, W. J. and MEIER, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory* **18** 394–412. https://doi.org/10.1080/09658211003702171

JARQUE, C. M. and BERA, A. K. (1987). A test for normality of observations and regression residuals. *Int. Stat. Rev.* **55** 163–172. MR0963337 https://doi.org/10.2307/1403192

JIN, Z., RISK, B. B. and MATTESON, D. S. (2019). Optimization and testing in linear non-Gaussian component analysis. *Stat. Anal. Data Min.* **12** 141–156. MR3957283 https://doi.org/10.1002/sam.11403

KAGAN, A. M., RAO, C. R. and LINNIK, Y. V. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

LERMAN-SINKOFF, D. B., SUI, J., RACHAKONDA, S., KANDALA, S., CALHOUN, V. D. and BARCH, D. M. (2017). Multimodal neural correlates of cognitive control in the Human Connectome Project. *NeuroImage* **163** 41–54. https://doi.org/10.1016/j.neuroimage.2017.08.081

LIU, J., PEARLSON, G., WINDEMUTH, A., RUANO, G., PERRONE-BIZZOZERO, N. I. and CALHOUN, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* **30** 241–255.

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 https://doi.org/10.1214/12-AOAS597

LOVE, M. (2019). Awesome multi-omics, https://github.com/mikelove/awesome-multi-omics.

MO, Q., WANG, S., SESHAN, V. E., OLSHEN, A. B., SCHULTZ, N., SANDER, C., POWERS, R. S., LADANYI, M. and SHEN, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA* **110** 4245–4250.

MOHANTY, R., SETHARES, W. A., NAIR, V. A. and PRABHAKARAN, V. (2020). Rethinking measures of functional connectivity via feature extraction. *Sci. Rep.* **10** 1–17.

NORDHAUSEN, K., OJA, H. and TYLER, D. E. (2016). Asymptotic and bootstrap tests for subspace dimension. Preprint. Available at arXiv:1611.04908.

NORDHAUSEN, K., OJA, H., TYLER, D. E. and VIRTA, J. (2017). Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. *IEEE Signal Process. Lett.* **24** 887–891.

OUYANG, X., CHEN, K., YAO, L., HU, B., WU, X., YE, Q., GUO, X., INITIATIVE, A. D. N. et al. (2015). Simultaneous changes in gray matter volume and white matter fractional anisotropy in Alzheimer's disease revealed by multimodal CCA and joint ICA. *Neuroscience* **301** 553–562.

PAKRAVAN, M. and SHAMSOLLAHI, M. B. (2019). Extraction and automatic grouping of joint and individual sources in multisubject fMRI data using higher order cumulants. *IEEE J. Biomed. Health Inform.* **23** 744–757. https://doi.org/10.1109/JBHI.2018.2840085

RACHAKONDA, S., LIU, J. and CALHOUN, V. (2012). Fusion ICA toolbox (FIT) manual. Albuquerque, NM: The MIND Research Network, University of New Mexico.

RISK, B. B. and GAYNANOVA, I. (2021). Supplement to "Simultaneous Non-Gaussian Component Analysis (SING) for Data Integration in Neuroimaging." https://doi.org/10.1214/21-AOAS1466SUPPA, https://doi.org/10.1214/21-AOAS1466SUPPB

RISK, B. B., MATTESON, D. S. and RUPPERT, D. (2019). Linear non-Gaussian component analysis via maximum likelihood. *J. Amer. Statist. Assoc.* **114** 332–343. MR3941258 https://doi.org/10.1080/01621459.2017.1407772

RISK, B. B., MATTESON, D. S., RUPPERT, D., ELOYAN, A. and CAFFO, B. S. (2014). An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics* **70** 224–236. MR3251683 https://doi.org/10.1111/biom.12111

SMITH, S. M., BECKMANN, C. F., ANDERSSON, J., AUERBACH, E. J., BIJSTERBOSCH, J., DOUAUD, G., DUFF, E., FEINBERG, D. A., GRIFFANTI, L. et al. (2013). Resting-state fMRI in the human connectome project. *NeuroImage* **80** 144–168.

SUI, J., ADALI, T., PEARLSON, G., YANG, H., SPONHEIM, S. R., WHITE, T. and CALHOUN, V. D. (2010). A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. *NeuroImage* **51** 123–134.

SUI, J., PEARLSON, G., CAPRIHAN, A., ADALI, T., KIEHL, K. A., LIU, J., YAMAMOTO, J. and CALHOUN, V. D. (2011). Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. *NeuroImage* **57** 839–855.

SUI, J., ADALI, T., YU, Q., CHEN, J. and CALHOUN, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods* **204** 68–81.

SUI, J., HE, H., PEARLSON, G. D., ADALI, T., KIEHL, K. A., YU, Q., CLARK, V. P., CASTRO, E., WHITE, T. et al. (2013). Three-way (N-way) fusion of brain imaging data based on mCCA+ jICA and its application to discriminating schizophrenia. *NeuroImage* **66** 119–132.

TANG, F., YANG, H., LI, L., JI, E., FU, Z. and ZHANG, Z. (2020). Fusion analysis of gray matter and white matter in bipolar disorder by multimodal CCA-joint ICA. *J. Affective Disorders* **263** 80–88.

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

VERGARA, V. M., ULLOA, A., CALHOUN, V. D., BOUTTE, D., CHEN, J. and LIU, J. (2014). A three-way parallel ICA approach to analyze links among genetics, brain structure and brain function. *NeuroImage* **98** 386–394.

VIRTA, J., NORDHAUSEN, K. and OJA, H. (2016). Projection pursuit for non-Gaussian independent components. Preprint. Available at arXiv:1612.05445.

WEN, Z. and YIN, W. (2013). A feasible method for optimization with orthogonality constraints. *Math. Program.* **142** 397–434. MR3127080 https://doi.org/10.1007/s10107-012-0584-1

WILLETTE, A. A., CALHOUN, V. D., EGAN, J. M., KAPOGIANNIS, D., INITIATIVE, A. D. N. et al. (2014). Prognostic classification of mild cognitive impairment and Alzheimer's disease: MRI independent component analysis. *Psychiatry Research. Neuroimaging* **224** 81–88.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

ZHOU, G., CICHOCKI, A., ZHANG, Y. and MANDIC, D. P. (2016a). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **27** 2426–2439. MR3571617 https://doi.org/10.1109/TNNLS.2015.2487364

ZHOU, G., ZHAO, Q., ZHANG, Y., ADALI, T., XIE, S. and CICHOCKI, A. (2016b). Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proc. IEEE* **104** 310–331.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1455–1477
https://doi.org/10.1214/21-AOAS1475

TENSOR QUANTILE REGRESSION WITH APPLICATION TO ASSOCIATION BETWEEN NEUROIMAGES AND HUMAN INTELLIGENCE

BY CAI LI* AND HEPING ZHANG†

*Department of Biostatistics, Yale University, *cai.li@yale.edu; †heping.zhang@yale.edu*

Human intelligence is usually measured by well-established psychometric tests through a series of problem solving. The recorded cognitive scores are continuous but usually heavy-tailed with potential outliers and violating the normality assumption. Meanwhile, magnetic resonance imaging (MRI) provides an unparalleled opportunity to study brain structures and cognitive ability. Motivated by association studies between MRI images and human intelligence, we propose a tensor quantile regression model, which is a general and robust alternative to the commonly used scalar-on-image linear regression. Moreover, we take into account rich spatial information of brain structures, incorporating low-rankness and piecewise smoothness of imaging coefficients into a regularized regression framework. We formulate the optimization problem as a sequence of penalized quantile regressions with a generalized Lasso penalty, based on tensor decomposition, and develop a computationally efficient alternating direction method of multipliers algorithm (ADMM) to estimate the model components. Extensive numerical studies are conducted to examine the empirical performance of the proposed method and its competitors. Finally, we apply the proposed method to a large-scale important dataset—the Human Connectome Project. We find that the tensor quantile regression can serve as a prognostic tool to assess future risk of cognitive impairment progression. More importantly, with the proposed method we are able to identify the most activated brain subregions associated with quantiles of human intelligence. The prefrontal and anterior cingulate cortex are found to be mostly associated with lower and upper quantile of fluid intelligence. The insular cortex associated with median of fluid intelligence is a rarely reported region.

REFERENCES

ASHBURNER, J. and FRISTON, K. J. (2000). Voxel-based morphometry—the methods. *NeuroImage* **11** 805–821.

ASHBURNER, J. and FRISTON, K. J. (2001). Why voxel-based morphometry should be used. *NeuroImage* **14** 1238–1243.

BILKER, W. B., HANSEN, J. A., BRENSINGER, C. M., RICHARD, J., GUR, R. E. and GUR, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment* **19** 354–369.

BONDELL, H. D., REICH, B. J. and WANG, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97** 825–838. MR2746154 https://doi.org/10.1093/biomet/asq048

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

BRANTLEY, H. L., GUINNESS, J. and CHI, E. C. (2020). Baseline drift estimation for air quality data using quantile trend filtering. *Ann. Appl. Stat.* **14** 585–604. MR4117821 https://doi.org/10.1214/19-AOAS1318

BRIOLLAIS, L. and DURRIEU, G. (2014). Application of quantile regression to recent genetic and-omic studies. *Human Genetics* **133** 951–966.

CHAN, E., MACPHERSON, S. E., BOZZALI, M., SHALLICE, T. and CIPOLOTTI, L. (2018). The influence of fluid intelligence, executive functions and premorbid intelligence on memory in frontal patients. *Front. Psychol.* **9** 926. https://doi.org/10.3389/fpsyg.2018.00926

DEARY, I. J. (2000). *Looking down on Human Intelligence*: *From Psychometrics to the Brain*. Oxford University Press, London.

Key words and phrases. Brain imaging, conditional quantile, fluid intelligence, generalized Lasso regularization, piecewise smoothness, tensor regression.

FENG, L., BI, X. and ZHANG, H. (2021). Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation. *J. Amer. Statist. Assoc.* **116** 144–158. MR4227681 https://doi.org/10.1080/01621459.2020.1766468

GLASSER, M. F., SOTIROPOULOS, S. N., WILSON, J. A., COALSON, T. S., FISCHL, B., ANDERSSON, J. L., XU, J., JBABDI, S., WEBSTER, M. et al. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80** 105–124.

GONG, Q.-Y., SLUMING, V., MAYES, A., KELLER, S., BARRICK, T., CEZAYIRLI, E. and ROBERTS, N. (2005). Voxel-based morphometry and stereology provide convergent evidence of the importance of medial prefrontal cortex for fluid intelligence in healthy adults. *NeuroImage* **25** 1175–1186.

GOOD, C. D., JOHNSRUDE, I. S., ASHBURNER, J., HENSON, R. N., FRISTON, K. J. and FRACKOWIAK, R. S. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* **14** 21–36. https://doi.org/10.1006/nimg.2001.0786

GU, Y., FAN, J., KONG, L., MA, S. and ZOU, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics* **60** 319–331. MR3847169 https://doi.org/10.1080/00401706.2017.1345703

HAIER, R. J., JUNG, R. E., YEO, R. A., HEAD, K. and ALKIRE, M. T. (2004). Structural brain variation and general intelligence. *NeuroImage* **23** 425–433.

HE, X., NG, P. and PORTNOY, S. (1998). Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 537–550. MR1625950 https://doi.org/10.1111/1467-9868.00138

HIDESE, S., OTA, M., MATSUO, J., ISHIDA, I., HIRAISHI, M., YOKOTA, Y., HATTORI, K., YOMOGIDA, Y. and KUNUGI, H. (2020). Correlation between the wechsler adult intelligence scale-metrics and brain structure in healthy individuals: A whole-brain magnetic resonance imaging study. *Front. Human Neurosci.* **14** 211.

JIANG, L., WANG, H. J. and BONDELL, H. D. (2013). Interquantile shrinkage in regression models. *J. Comput. Graph. Statist.* **22** 970–986. MR3173752 https://doi.org/10.1080/10618600.2012.707454

KIEVIT, R. A., DAVIS, S. W., MITCHELL, D. J., TAYLOR, J. R., DUNCAN, J. and HENSON, R. N. (2014). Distinct aspects of frontal lobe structure mediate age-related differences in fluid intelligence and multitasking. *Nat. Commun.* **5** 1–10.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657 https://doi.org/10.1017/CBO9780511754098

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417 https://doi.org/10.1093/biomet/81.4.673

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

LI, C. and ZHANG, H. (2021). Supplement to "Tensor quantile regression with application to association between neuroimages and human intelligence." https://doi.org/10.1214/21-AOAS1475SUPPA, https://doi.org/10.1214/21-AOAS1475SUPPB

LI, Z., SUK, H.-I., SHEN, D. and LI, L. (2016). Sparse multi-response tensor regression for Alzheimer's disease study with multivariate clinical assessments. *IEEE Trans. Med. Imag.* **35** 1927–1936.

LI, X., XU, D., ZHOU, H. and LI, L. (2018). Tucker tensor regression and neuroimaging analysis. *Stat. Biosci.* **10** 520–545.

LIU, M., ZHANG, D. and SHEN, D. (2014). Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* **35** 1305–1319.

LUDERS, E., GASER, C., JANCKE, L. and SCHLAUG, G. (2004). A voxel-based approach to gray matter asymmetries. *NeuroImage* **22** 656–664.

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

MUGLER, J. P. and BROOKEMAN, J. R. (1990). Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). *Magn. Reson. Med.* **15** 152–157.

REINHOLD, J. C., DEWEY, B. E., CARASS, A. and PRINCE, J. L. (2019). Evaluating the impact of intensity normalization on MR image synthesis. In *Medical Imaging* 2019: *Image Processing* **10949** 109493H. International Society for Optics and Photonics.

RISCH, N. and ZHANG, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268** 1584–1589.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms *Phys. D, Nonlinear Phenom.* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

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.

SHINOHARA, R. T., SWEENEY, E. M., GOLDSMITH, J., SHIEE, N., MATEEN, F. J., CALABRESI, P. A., JARSO, S., PHAM, D. L., REICH, D. S. and CRAINICEANU, C. M. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage Clin.* **6** 9–19.

SUN, W. W. and LI, L. (2017). STORE: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** Paper No. 135, 37. MR3763769

THE MATHWORKS INC. (2020). MATLAB, Version 9.8.0.1538580 (R2020a).

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205 https://doi.org/10.1214/11-AOS878

TOGA, A. W. and THOMPSON, P. M. (2005). Genetics of brain structure and intelligence. *Annu. Rev. Neurosci.* **28** 1–23. https://doi.org/10.1146/annurev.neuro.28.061604.135655

UĞURBIL, K., XU, J., AUERBACH, E. J., MOELLER, S., VU, A. T., DUARTE-CARVAJALINO, J. M., LENGLET, C., WU, X., SCHMITTER, S. et al. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project. *NeuroImage* **80** 80–104.

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

VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E., YACOUB, E., UGURBIL, K., CONSORTIUM, W.-M. H. et al. (2013). The WU-Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

WANG, X. and ZHU, H. (2017). Generalized scalar-on-image regression models via total variation. *J. Amer. Statist. Assoc.* **112** 1156–1168. MR3735367 https://doi.org/10.1080/01621459.2016.1194846

WATKINS, K. E., PAUS, T., LERCH, J. P., ZIJDENBOS, A., COLLINS, D. L., NEELIN, P., TAYLOR, J., WORSLEY, K. J. and EVANS, A. C. (2001). Structural asymmetries in the human brain: A voxel-based statistical analysis of 142 MRI scans. *Cereb. Cortex* **11** 868–877.

WEI, Y., PERE, A., KOENKER, R. and HE, X. (2006). Quantile regression methods for reference growth charts. *Stat. Med.* **25** 1369–1382. MR2226792 https://doi.org/10.1002/sim.2271

WINKLER, A. M., KOCHUNOV, P., BLANGERO, J., ALMASY, L., ZILLES, K., FOX, P. T., DUGGIRALA, R. and GLAHN, D. C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage* **53** 1135–1146.

YU, L. and LIN, N. (2017). ADMM for penalized quantile regression in big data. *Int. Stat. Rev.* **85** 494–518. MR3723614 https://doi.org/10.1111/insr.12221

YUAN, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **50** 813–829. MR2207010 https://doi.org/10.1016/j.csda.2004.10.008

ZHANG, X., LI, L., ZHOU, H., ZHOU, Y., SHEN, D. and ADNI (2019). Tensor generalized estimating equations for longitudinal imaging analysis. *Statist. Sinica* **29** 1977–2005. MR3970344

ZHOU, H. and LI, L. (2014). Regularized matrix regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 463–483. MR3164874 https://doi.org/10.1111/rssb.12031

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1478–1498
https://doi.org/10.1214/21-AOAS1473

DIAGNOSIS-GROUP-SPECIFIC TRANSITIONAL CARE PROGRAM RECOMMENDATIONS FOR 30-DAY REHOSPITALIZATION REDUCTION

BY MENGGANG YU[1], CHENSHENG KUANG[2], JARED D. HULING[3] AND MAUREEN SMITH[4]

[1]*Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, meyu@biostat.wisc.edu*

[2]*Department of Statistics, University of Wisconsin-Madison, chenshengkuang@gmail.com*

[3]*Division of Biostatistics, School of Public Health, University of Minnesota, huling@umn.edu*

[4]*Department of Population Health Sciences, Department of Family Medicine & Community Health, and Health Innovation Program, University of Wisconsin-Madison, maureensmith@wisc.edu*

Thirty-day rehospitalization rate is a well-studied and important measure reflecting the overall performance of health systems. Recently, transitional care (TC) programs have been initiated to reduce avoidable rehospitalizations. These programs typically ask nurses to follow-up with patients after the hospitalization to manage issues and reduce the risk of rehospitalizations during health care transitions. As rehospitalization is a complex process that depends on many factors, it is unlikely that these interventions are effective for all patients across a diverse population. In this paper we consider individualized intervention or treatment recommendation rules (ITRs) aimed at maximizing overall treatment effectiveness. We investigate our approach in a setting where patients are divided into two diagnosis related groups, medically complicated and uncomplicated. As the treatment effects can greatly vary between the two groups, we allow our recommendation rules to be group specific. In particular, our approach can accommodate scale differences in treatment effects and utilize a tuning parameter to drive the similarity of the estimated ITRs between groups. Computation is achieved by transforming our problem into a form solvable by existing software, and a wrapper R package is developed for our proposed treatment recommendation framework. We conduct extensive evaluation through both simulation studies and analysis of a TC program.

REFERENCES

BETANCOURT, J. R., TAN-MCGRORY, A. and KENST, K. (2015). Guide to preventing readmissions among racially and ethnically diverse Medicare beneficiaries. *Health*.

BRADLEY, E. H., CURRY, L., HORWITZ, L. I., SIPSMA, H., THOMPSON, J. W., ELMA, M., WALSH, M. N. and KRUMHOLZ, H. M. (2012). Contemporary evidence about hospital strategies for reducing 30-day readmissions: A national study. *J. Am. Coll. Cardiol.* **60** 607–614.

BRADLEY, E. H., CURRY, L., HORWITZ, L. I., SIPSMA, H., WANG, Y., WALSH, M. N., GOLDMANN, D., WHITE, N., PIÑA, I. L. et al. (2013). Hospital strategies associated with 30-day readmission rates for patients with heart failure. *Circulation*: *Cardiovascular Quality and Outcomes* **6** 444–450.

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534 https://doi.org/10.1111/biom.12676

CLOONAN, P., WOOD, J. and RILEY, J. B. (2013). Reducing 30-day readmissions: Health literacy strategies. *Journal of Nursing Administration* **43** 382–387.

COLEMAN, E. A., PARRY, C., CHALMERS, S. and MIN, S.-J. (2006). The care transitions intervention: Results of a randomized controlled trial. *Arch. Intern. Med.* **166** 1822–1828.

COX, D. R. (1958). *Planning of Experiments*. *A Wiley Publication in Applied Statistics*. Wiley, New York. MR0095561

DONZÉ, J., AUJESKY, D., WILLIAMS, D. and SCHNIPPER, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Internal Medicine* **173** 632–638.

Key words and phrases. Heterogeneity of treatment effect, observational data, rehospitalization, subgroup identification, data integration.

EVASHWICK, C. (2005). *The Continuum of Long-Term Care*. Cengage Learning.
FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273
FOX, T., BRUMMIT, P. S., FERGUSON-WOLF, M., ABERNETHY, M. et al. (2000). Position of the American Dietetic Association: Nutrition, aging, and the continuum of care. *Journal of the Academy of Nutrition and Dietetics* **100** 580.
HANSEN, L. O., YOUNG, R. S., HINAMI, K., LEUNG, A. and WILLIAMS, M. V. (2011). Interventions to reduce 30-day rehospitalization: A systematic review. *Ann. Intern. Med.* **155** 520–528.
HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. With discussion and a reply by the author. MR0867618
HULING, J. D. and CHIEN, P. (2021). Fast penalized regression and cross validation for tall data with the oem package. *J. Stat. Softw.* To appear.
IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470. MR3086426 https://doi.org/10.1214/12-AOAS593
JENCKS, S. F., WILLIAMS, M. V. and COLEMAN, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N. Engl. J. Med.* **360** 1418–1428.
KEHL, V. and ULM, K. (2006). Responder identification in clinical trials with censored data. *Comput. Statist. Data Anal.* **50** 1338–1355. MR2224375 https://doi.org/10.1016/j.csda.2004.11.015
KIND, A. J., BRENNY-FITZPATRICK, M., LEAHY-GROSS, K., MIRR, J., CHAPMAN, E., FREY, B. and HOULAHAN, B. (2016). Harnessing protocolized adaptation in dissemination: Successful implementation and sustainment of the veterans affairs coordinated-transitional care program in a non-veterans affairs hospital. *J. Amer. Geriatr. Soc.* **64** 409–416.
KRIPALANI, S., THEOBALD, C. N., ANCTIL, B. and VASILEVSKIS, E. E. (2014). Reducing hospital readmission rates: Current strategies and future directions. *Annual Review of Medicine* **65** 471–485.
LEPPIN, A. L., GIONFRIDDO, M. R., KESSLER, M., BRITO, J. P., MAIR, F. S., GALLACHER, K., WANG, Z., ERWIN, P. J., SYLVESTER, T. et al. (2014). Preventing 30-day hospital readmissions: A systematic review and meta-analysis of randomized trials. *JAMA Internal Medicine* **174** 1095–1107.
LIPKOVICH, I., DMITRIENKO, A. and D'AGOSTINO, R. B. SR. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* **36** 136–196. MR3580950 https://doi.org/10.1002/sim.7064
MCILVENNAN, C. K., EAPEN, Z. J. and ALLEN, L. A. (2015). Hospital readmissions reduction program. *Circulation* **131** 1796–1803.
NAYLOR, M. D., AIKEN, L. H., KURTZMAN, E. T., OLDS, D. M. and HIRSCHMAN, K. B. (2011). The importance of transitional care in achieving health reform. *Health Aff.* **30** 746–754.
NORRVING, B. and KISSELA, B. (2013). The global burden of stroke and need for a continuum of care. *Neurology* **80** S5–S12.
OLLIER, E. and VIALLON, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika* **104** 83–96. MR3626476 https://doi.org/10.1093/biomet/asw065
QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864
RAU, J. (2014). Medicare fines 2610 hospitals in third round of readmission penalties. *Kaiser Health News* **2**.
ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730
RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880
SHI, C., SONG, R., LU, W. and FU, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 681–702. MR3849339 https://doi.org/10.1111/rssb.12273
SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013a). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250
SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013b). SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization. R package version 1.1.
STEVENS, S. (2015). Preventing 30-day readmissions. *Nursing Clinics* **50** 123–137.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
XIONG, S., DAI, B., HULING, J. and QIAN, P. Z. G. (2016). Orthogonalizing EM: A design-based least squares algorithm. *Technometrics* **58** 285–293. MR3520658 https://doi.org/10.1080/00401706.2015.1054436
YU, M., KUANG, C., HULING, J. D and SMITH, M. (2021). Supplement to "Diagnosis-group-specific transitional care program recommendations for 30-day rehospitalization reduction." https://doi.org/10.1214/21-AOAS1473SUPPA, https://doi.org/10.1214/21-AOAS1473SUPPB

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114. MR4027418 https://doi.org/10.1002/sta.411

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1499–1528
https://doi.org/10.1214/20-AOAS1436

GLOBAL ESTIMATION AND SCENARIO-BASED PROJECTIONS OF SEX RATIO AT BIRTH AND MISSING FEMALE BIRTHS USING A BAYESIAN HIERARCHICAL TIME SERIES MIXTURE MODEL

BY FENGQING CHAO[1], PATRICK GERLAND[2], ALEX R. COOK[3] AND LEONTINE ALKEMA[4]

[1]*Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, fengqing.chao@kaust.edu.sa*

[2]*United Nations Population Division, DESA, United Nations, gerland@un.org*

[3]*Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, ephcar@nus.edu.sg*

[4]*School of Public Health and Health Sciences, University of Massachusetts, Amherst, lalkema@umass.edu*

The sex ratio at birth (SRB) is defined as the ratio of male to female live births. The SRB imbalance in parts of the world over the past several decades is a direct consequence of sex-selective abortion, driven by the coexistence of son preference, readily available technology of prenatal sex determination and fertility decline. Estimation and projection of the degree of SRB imbalance is complicated because of variability in SRB reference levels and because of the uncertainty associated with SRB observations.

We develop Bayesian hierarchical time series mixture models for SRB estimation and scenario-based projections for all countries from 1950 to 2100. We model the SRB regional and national reference levels and the fluctuation around national reference levels. We identify countries at risk of SRB imbalances and model both: (i) the absence or presence of sex ratio transitions in such countries and, if present, (ii) the transition process. The transition model of SRB imbalance captures three stages (increase, stagnation and convergence back to SRB baselines). The model identifies countries with statistical evidence of SRB inflation in a fully Bayesian approach. The scenario-based SRB projections are based on the sex ratio transition model with varying assumptions regarding the occurrence of a sex ratio transition in at-risk countries. Projections are used to quantify the future burden of missing female births due to sex-selective abortions under different scenarios.

REFERENCES

ALKEMA, L., WONG, M. B. and SEAH, P. R. (2012). Monitoring progress towards Millennium Development Goal 4: A call for improved validation of under-five mortality rate estimates. *Statistics*, *Politics*, *and Policy* **3**.

ALKEMA, L., CHAO, F., YOU, D., PEDERSEN, J. and SAWYER, C. C. (2014). National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: A systematic assessment. *Lancet Glob Health* **2** e521–e530. https://doi.org/10.1016/S2214-109X(14)70280-3

ALLAHBADIA, G. N. (2002). The 50 million missing women. *Journal of Assisted Reproduction and Genetics* **19** 411–416.

ATTANÉ, I. and GUILMOTO, C. Z. (2007). *Watering the Neighbour's Garden*: *The Growing Demographic Female Deficit in Asia*. Committee for International Cooperation in National Research in Demography, Paris.

BASTEN, S. and VERROPOULOU, G. (2013). Maternity migration and the increased sex ratio at birth in Hong Kong SAR. *Popul. Stud.* **67** 323–334.

BONGAARTS, J. (2013). The implementation of preferences for male offspring. *Popul. Dev. Rev.* **39** 185–208.

BONGAARTS, J. and GUILMOTO, C. Z. (2015). How many more missing women? Excess female mortality and prenatal sex selection, 1970–2050. *Popul. Dev. Rev.* **41** 241–269.

CATALANO, R., BRUCKNER, T., MARKS, A. R. and ESKENAZI, B. (2006). Exogenous shocks to the human sex ratio: The case of September 11, 2001 in New York city. *Human Reproduction* **21** 3127–3131.

Key words and phrases. Bayesian hierarchical model, probabilistic scenario-based projection, time series analysis, sex-selective abortion, sex ratio transition, missing female births.

CHAHNAZARIAN, A. (1988). Determinants of the sex ratio at birth: Review of recent literature. *Social Biology* **35** 214–235.

CHAO, F., KC, S. and OMBAO, H. (2020). Levels and trends in the sex ratio at birth in seven provinces of Nepal between 1980 and 2016 with probabilistic projections to 2050: A Bayesian modeling approach. Preprint. Available at arXiv:2007.00437.

CHAO, F. and YADAV, A. K. (2019). Levels and trends in the sex ratio at birth and missing female births for 29 states and union territories in India 1990–2016: A Bayesian modeling study. *Foundations of Data Science* **1** 177–196.

CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019a). Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. *Proc. Natl. Acad. Sci. USA* **116** 9303–9311.

CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019b). SRB database (pnas.1812593116.sd01). Available at https://www.pnas.org/highwire/filestream/859048/field_highwire_adjunct_files/1/pnas.1812593116.sd01.xlsx.

CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2019c). Web appendix systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. Available at https://www.pnas.org/content/pnas/suppl/2019/04/10/1812593116.DCSupplemental/pnas.1812593116.sapp.pdf. https://doi.org/10.6084/m9.figshare.12442373

CHAO, F., GUILMOTO, C. Z., KC, S. and OMBAO, H. (2020). Probabilistic projection of the sex ratio at birth and missing female births by state and union territory in India. *PLoS ONE* **15** e0236673.

CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2021a). Supplement to "Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model." https://doi.org/10.1214/20-AOAS1436SUPPA

CHAO, F., GERLAND, P., COOK, A. R. and ALKEMA, L. (2021b). Supplement to "Global estimation and scenario-based projections of sex ratio at birth and missing female births using a Bayesian hierarchical time series mixture model." https://doi.org/10.1214/20-AOAS1436SUPPB

CHEN, C., CHOU, S.-Y., GIMENEZ, L. and LIU, J.-T. (2020). The quantity of education and preference for sons: Evidence from Taiwan's compulsory education reform. *China Econ. Rev.* **59** 101369.

CHOI, E. J. and HWANG, J. (2020). Transition of son preference: Evidence from South Korea. *Demography* 1–26.

DAS GUPTA, M., ZHENGHUA, J., BOHUA, L., ZHENMING, X., CHUNG, W. and HWA-OK, B. (2003). Why is son preference so persistent in East and South Asia? A cross-country study of China, India and the Republic of Korea. *The Journal of Development Studies* **40** 153–187.

DRÉZE, J. and SEN, A. (1990). *Hunger and Public Action*. Clarendon Press, Oxford.

DUBUC, S. and COLEMAN, D. (2007). An increase in the sex ratio of births to India-born mothers in England and Wales: Evidence for sex-selective abortion. *Popul. Dev. Rev.* **33** 383–400.

DUTHÉ, G., MESLÉ, F., VALLIN, J., BADURASHVILI, I. and KUYUMJYAN, K. (2012). High sex ratios at birth in the Caucasus: Modern technology to satisfy old desires. *Popul. Dev. Rev.* **38** 487–501.

FUKUDA, M., FUKUDA, K., SHIMIZU, T. and MØLLER, H. (1998). Decline in sex ratio at birth after Kobe earthquake. *Human Reproduction* **13** 2321–2322.

GARENNE, M. (2002). Sex ratios at birth in African populations: A review of survey data. *Hum. Biol.* **74** 889–900. https://doi.org/10.1353/hub.2003.0003

GARENNE, M. (2008). Poisson variations of the sex ratio at birth in African demographic surveys. *Hum. Biol.* **80** 473–482.

GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–511.

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

GEORGE, S. M. (2002). Sex selection/determination in India: Contemporary developments. *Reproductive Health Matters* **10** 190–192.

GOODKIND, D. (1997). *Sex-Selective Abortion*, *Reproductive Rights*, *and the Greater Locus of Gender Discrimination in Family Formation*: *Cairo's Unresolved Questions*. Univ. Michigan, Population Studies Center.

GOODKIND, D. (2011). Child underreporting, fertility, and sex ratio imbalance in China. *Demography* **48** 291–316. https://doi.org/10.1007/s13524-010-0007-y

GRAFFELMAN, J. and HOEKSTRA, R. F. (2000). A statistical analysis of the effect of warfare on the human secondary sex ratio. *Hum. Biol.* 433–445.

GUILMOTO, C. Z. (2009). The sex ratio transition in Asia. *Popul. Dev. Rev.* **35** 519–549.

GUILMOTO, C. Z. (2012a). Sex imbalances at birth: Trends, consequences and policy implications. UNFPA, United Nations Population Fund of Asia and the Pacific Regional Office, Thailand.

GUILMOTO, C. Z. (2012b). Skewed sex ratios at birth and future marriage squeeze in China and India, 2005–2100. *Demography* **49** 77–100. https://doi.org/10.1007/s13524-011-0083-7
GUILMOTO, C. Z. (2012c). Son preference, sex selection, and kinship in Vietnam. *Popul. Dev. Rev.* **38** 31–54. https://doi.org/10.1111/j.1728-4457.2012.00471.x
GUILMOTO, C. Z. (2015). Mapping the diversity of gender preferences and sex imbalances in Indonesia in 2010. *Popul. Stud.* (*Camb.*) **69** 299–315. https://doi.org/10.1080/00324728.2015.1091603
GUILMOTO, C. Z., CHAO, F. and KULKARNI, P. M. (2020). On the estimation of female births missing due to prenatal sex selection. *Popul. Stud.* (*Camb.*) **74** 283–289. https://doi.org/10.1080/00324728.2020.1762912
GUILMOTO, C. Z., HOÀNG, X. and VAN, T. N. (2009). Recent increase in sex ratio at birth in viet nam. *PLoS ONE* **4** e4624.
GUILMOTO, C. Z. and REN, Q. (2011). Socio-economic differentials in birth masculinity in China. *Development and Change* **42** 1269–1296.
GUPTA, M. D., CHUNG, W. and SHUZHUO, L. (2009). Evidence for an incipient decline in numbers of missing girls in China and India. *Popul. Dev. Rev.* **35** 401–416.
HUDSON, V. M. and DEN BOER, A. (2004). *Bare Branches*: *The Security Implications of Asia's Surplus Male Population*. MIT Press, Cambridge.
JAMES, W. H. (1984). The sex ratios of black births. *Ann. Hum. Biol.* **11** 39–44. https://doi.org/10.1080/03014468400006871
JAMES, W. H. (1985). The sex ratio of oriental births. *Ann. Hum. Biol.* **12** 485–487.
JAMES, W. H. (1987). The human sex ratio. Part 1: A review of the literature. *Hum. Biol.* 721–752.
JIANG, Q., GE, T. and TAI, X. (2019). Change in China's sex ratio at birth since 2000: A decomposition at the provincial level. *Appl. Spat. Anal. Policy* 1–28.
KABA, A. J. (2008). Sex ratio at birth and racial differences: Why do black women give birth to more females than non-black women? *Afr. J. Reprod. Health* **12** 139–150.
LIN, T.-C. (2009). The decline of son preference and rise of gender indifference in Taiwan since 1990. *Demogr. Res.* **20** 377–402. https://doi.org/10.4054/DemRes.2009.20.16
LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34** 1–41. MR0415861
MADAN, K. and BREUNING, M. H. (2013). Impact of prenatal technologies on the sex ratio in India: An overview. *Genetics in Medicine* **16** 425–432.
MARCUS, M., KIELY, J., XU, F., MCGEEHIN, M., JACKSON, R. and SINKS, T. (1998). Changing sex ratio in the United States, 1969–1995. *Fertility and Sterility* **70** 270–273.
MARTIN, A. D., QUINN, K. M. and PARK, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *J. Stat. Softw.* **42** 22.
MATHEWS, T. J. and HAMILTON, B. E. (2005). Trend analysis of the sex ratio at birth in the United States. *Natl. Vital Stat. Rep.* **53** 1–17.
MESLÉ, F., VALLIN, J. and BADURASHVILI, I. (2007). A sharp increase in sex ratio at birth in the Caucasus. Why? How? *Watering the Neighbour's Garden*: *The Growing Demographic Female Deficit in Asia*, *Paris*: *Committee for International Cooperation in National Research in Demography* 73–88.
OOMMAN, N. and GANATRA, B. R. (2002). Sex selection: The systematic elimination of girls. *Reproductive Health Matters* **10** 184–188.
PARK, C. B. and CHO, N.-H. (1995). Consequences of son preference in a low-fertility society: Imbalance of the sex ratio at birth in Korea. *Popul. Dev. Rev.* 59–84.
PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria.
PLUMMER, M. (2011). rjags: Bayesian graphical models using MCMC. R package version 3-5.
PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.
R CORE TEAM (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
RUDER, A. (1985). Paternal-age and birth-order effect on the human secondary sex ratio. *Am. J. Hum. Genet.* **37** 362–372.
ŠEVČÍKOVÁ, H., ALKEMA, L. and RAFTERY, A. E. (2011). bayesTFR: An R package for probabilistic projections of the total fertility rate. *J. Stat. Softw.* **43** 1–29.
ŠEVČÍKOVÁ, H., ALKEMA, L., RAFTERY, A. E., FOSDICK, B. and GERLAND, P. (2019). bayesTFR: Bayesian fertility projection. R package and documentation version 6.4-0. R-CRAN. Available at https://cran.r-project.org/web/packages/bayesTFR/. Accessed 16 Feb 2020.
SONG, S. (2012). Does famine influence sex ratio at birth? Evidence from the 1959–1961 great leap forward famine in China. *Proc. Roy. Soc., Biol. Sci.* **279** 2883–2890.

SU, Y. S. and YAJIMA, M. (2011). R2jags: A package for running jags from R. R package version 0.02-17.
TAFURO, S. and GUILMOTO, C. Z. (2020). Skewed sex ratios at birth: A review of global trends. *Early Hum. Dev.* **141** 104868. https://doi.org/10.1016/j.earlhumdev.2019.104868
TANDON, S. L. and SHARMA, R. (2006). Female foeticide and infanticide in India: An analysis of crimes against girl children. *International Journal of Criminal Justice Sciences* **1**.
UNITED NATIONS, DESA (2019). World population prospects: The 2019 revision. Available at http://esa.un.org/unpd/wpp/Download/Standard/Population/. Accessed 16 Feb 2020.
VEHTARI, A., GELMAN, A., YAO, Y. and GABRY, J. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.1.0 1003.
VENERO FERNÁNDEZ, S. J., MEDINA, R. S., BRITTON, J. and FOGARTY, A. W. (2011). The association between living through a prolonged economic depression and the male: Female birth ratio–a longitudinal study from Cuba, 1960–2008. *Am. J. Epidemiol*. **174** 1327–1331.
VISARIA, P. M. (1967). Sex ratio at birth in territories with a relatively complete registration. *Eugenics Quarterly* **14** 132–142.
VOLLSET, S. E., GOREN, E., YUAN, C.-W., CAO, J., SMITH, A. E., HSIAO, T., BISIGNANO, C., AZHAR, G. S., CASTRO, E. et al. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: A forecasting analysis for the global burden of disease study. *Lancet*.
VU, T. M. and YAMADA, H. (2020). Sex ratio and religion in Vietnam. *Munich Personal RePEc Archive*.
YAO, Y., VEHTARI, A. and GELMAN, A. (2020). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. Preprint. Available at arXiv:2006.12335.

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1529–1555
https://doi.org/10.1214/21-AOAS1439

PARTIAL-MASTERY COGNITIVE DIAGNOSIS MODELS

BY ZHUORAN SHANG[1], ELENA A. EROSHEVA[2] AND GONGJUN XU[3]

[1]*School of Statistics, University of Minnesota, zhuoranshang@gmail.com*

[2]*Department of Statistics, School of Social Work, & the Center for Statistics and the Social Sciences, University of Washington, erosheva@uw.edu*

[3]*Department of Statistics, University of Michigan, gongjun@umich.edu*

Cognitive diagnosis models (CDMs) are a family of discrete latent attribute models that serve as statistical basis in educational and psychological cognitive diagnosis assessments. CDMs aim to achieve fine-grained inference on individuals' latent attributes, based on their observed responses to a set of designed diagnostic items. In the literature CDMs usually assume that items require mastery of specific latent attributes and that each attribute is either fully mastered or not mastered by a given subject. We propose a new class of models, partial mastery CDMs (PM-CDMs), that generalizes CDMs by allowing for partial mastery levels for each attribute of interest. We demonstrate that PM-CDMs can be represented as restricted latent class models. Relying on the latent class representation, we propose a Bayesian approach for estimation. We present simulation studies to demonstrate parameter recovery, to investigate the impact of model misspecification with respect to partial mastery and to develop diagnostic tools that could be used by practitioners to decide between CDMs and PM-CDMs. We use two examples of real test data—the fraction subtraction and the English tests—to demonstrate that employing PM-CDMs not only improves model fit, compared to CDMs, but also can make substantial difference in conclusions about attribute mastery. We conclude that PM-CDMs can lead to more effective remediation programs by providing detailed individual-level information about skills learned and skills that need to study.

REFERENCES

AIROLDI, E. M., BLEI, D. M., EROSHEVA, E. A. and FIENBERG, S. E. (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press, Boca Raton, FL.

BLEI, D. M. and LAFFERTY, J. D. (2007). A correlated topic model of *Science*. *Ann. Appl. Stat.* **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOLT, D. M. and LALL, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Appl. Psychol. Meas.* **27** 395–414. MR2005523 https://doi.org/10.1177/0146621603258350

CHEN, J. and DE LA TORRE, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Appl. Psychol. Meas.* **37** 419–437.

CHEN, Y., LIU, J., XU, G. and YING, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *J. Amer. Statist. Assoc.* **110** 850–866. MR3367269 https://doi.org/10.1080/01621459.2014.934827

CHEN, Y., CULPEPPER, S. A., CHEN, Y. and DOUGLAS, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika* **83** 89–108. MR3767014 https://doi.org/10.1007/s11336-017-9579-4

CHIU, C.-Y., DOUGLAS, J. A. and LI, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika* **74** 633–665. MR2565331 https://doi.org/10.1007/s11336-009-9125-0

CHIU, C.-Y., KÖHN, H.-F., ZHENG, Y. and HENSON, R. (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika* **81** 1069–1092. MR3583919 https://doi.org/10.1007/s11336-016-9534-9

CULPEPPER, S. A. (2015). Bayesian estimation of the dina model with Gibbs sampling. *J. Educ. Behav. Stat.* **40** 454–476.

Key words and phrases. Cognitive diagnosis, restricted latent class models, mixed membership models.

CULPEPPER, S. A. and CHEN, Y. (2019). Development and application of an exploratory reduced reparameterized unified model. *J. Educ. Behav. Stat.* **44** 3–24.
CULPEPPER, S. A. and HUDSON, A. (2018). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Appl. Psychol. Meas.* **42** 99–115.
DECARLO, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, class sizes, and the Q-matrix. *Appl. Psychol. Meas.* **35** 8–26.
DECARLO, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Appl. Psychol. Meas.* **36** 447–468.
DE LA TORRE, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *J. Educ. Meas.* **45** 343–362.
DE LA TORRE, J. (2009). DINA model and parameter estimation: A didactic. *J. Educ. Behav. Stat.* **34** 115–130.
DE LA TORRE, J. (2011). The generalized DINA model framework. *Psychometrika* **76** 179–199. MR2788881 https://doi.org/10.1007/s11336-011-9207-7
DE LA TORRE, J. and CHIU, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika* **81** 253–273. MR3505366 https://doi.org/10.1007/s11336-015-9467-8
DE LA TORRE, J. and DOUGLAS, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* **69** 333–353. MR2272454 https://doi.org/10.1007/BF02295640
DIBELLO, L. V., STOUT, W. F. and ROUSSOS, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively Diagnostic Assessment* (P. D. Nichols, S. F. Chipman and R. L. Brennan, eds.) 361–390. Erlbaum Associates, Hillsdale, NJ.
EMBRETSON, S. E. and YANG, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika* **78** 14–36. MR3042817 https://doi.org/10.1007/s11336-012-9296-y
EROSHEVA, E. A. (2004). Partial membership models with application to disability survey data. In *Statistical Data Mining and Knowledge Discovery* 117–134. CRC Press/CRC, Boca Raton, FL. MR2048951
EROSHEVA, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika* **70** 619–628. MR2272507 https://doi.org/10.1007/s11336-001-0899-y
EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745 https://doi.org/10.1214/07-AOAS126
EROSHEVA, E., FIENBERG, S. and LAFFERTY, J. (2004). Mixed-membership models of scientific publications. *Proc. Natl. Acad. Sci. USA* **101** 5220–5227.
FENG, Y., HABING, B. T. and HUEBNER, A. (2014). Parameter estimation of the reduced rum using the em algorithm. *Appl. Psychol. Meas.* **38** 137–150.
GALYARDT, A. (2015). Interpreting mixed membership models: Implications of Erosheva's representation theorem. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 39–65. CRC Press, Boca Raton, FL. MR3380024
GEORGE, A. C., ROBITZSCH, A., KIEFER, T., GROSS, J. and ÜNLÜ, A. (2016). The R package CDM for cognitive diagnosis models. *J. Stat. Softw.* **74** 1–24.
GU, Y. and XU, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* **84** 468–483. MR3947373 https://doi.org/10.1007/s11336-018-9619-8
GU, Y. and XU, G. (2020). Partial identifiability of restricted latent class models. *Ann. Statist.* **48** 2082–2107. MR4134787 https://doi.org/10.1214/19-AOS1878
GU, Y., LIU, J., XU, G. and YING, Z. (2018). Hypothesis testing of the *Q*-matrix. *Psychometrika* **83** 515–537. MR3851945 https://doi.org/10.1007/s11336-018-9629-6
HABERMAN, S. (1995). Book review of "Statistical applications using fuzzy sets," by K.G. Manton, M.A. Woodbury, and L.S. Corder. *J. Amer. Statist. Assoc.* **90** 1131–1133.
HANSEN, M., CAI, L., MONROE, S. and LI, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *Br. J. Math. Stat. Psychol.* **69** 225–252.
HARTZ, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities*: *Blending Theory with Practicality*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Univ. Illinois at Urbana-Champaign. MR2703174
HARTZ, S. and ROUSSOS, L. (2008). The fusion model for skills diagnosis: Blending theory with practicality. *ETS Res. Rep. Ser.* **2008** i–57.
HENSON, R. A., TEMPLIN, J. L. and WILLSE, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* **74** 191–210. MR2507377 https://doi.org/10.1007/s11336-008-9089-5
HONG, H., WANG, C., LIM, Y. S. and DOUGLAS, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Appl. Psychol. Meas.* **39** 31–43. https://doi.org/10.1177/0146621614524981

JUNKER, B. W. and SIJTSMA, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **25** 258–272. MR1842982 https://doi.org/10.1177/01466210122032064

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

KÖHN, H.-F. and CHIU, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika* **82** 112–132. MR3614810 https://doi.org/10.1007/s11336-016-9536-7

KÖHN, H.-F. and CHIU, C.-Y. (2018). How to build a complete Q-matrix for a cognitively diagnostic test. *J. Classification* **35** 273–299. MR3849135 https://doi.org/10.1007/s00357-018-9255-0

LIU, Y., DOUGLAS, J. A. and HENSON, R. A. (2007). Testing person fit in cognitive diagnosis. In *Annual Meeting of the National Council on Measurement in Education* (*NCME*), *Chicago*, *IL*, *April*.

LIU, J., XU, G. and YING, Z. (2012). Data-driven learning of Q-matrix. *Appl. Psychol. Meas.* **36** 548–564. https://doi.org/10.1177/0146621612456591

LIU, J., XU, G. and YING, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli* **19** 1790–1817. MR3129034 https://doi.org/10.3150/12-BEJ430

MA, W. and DE LA TORRE, J. (2019). Gdina: The generalized dina model framework. R package version, 2.3.2. Retrieved from https://CRAN.R-project.org/package=GDINA.

MANTON, K. G., WOODBURY, M. A. and TOLLEY, H. D. (1994). *Statistical Applications Using Fuzzy Sets*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1269319

RECKASE, M. (2009). *Multidimensional Item Response Theory* **150**. Springer, Berlin.

RUPP, A. A. and TEMPLIN, J. L. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educ. Psychol. Meas.* **68** 78–98. MR2416509 https://doi.org/10.1177/0013164407301545

RUPP, A. A., TEMPLIN, J. L. and HENSON, R. A. (2010). *Diagnostic Measurement*: *Theory*, *Methods*, *and Applications*. Guilford Press, New York.

SHANG, Z., EROSHEVA, E. A. and XU, G. (2021). Supplement to "Partial-mastery cognitive diagnosis models." https://doi.org/10.1214/21-AOAS1439SUPP

STOUT, W., HENSON, R., DIBELLO, L. and SHEAR, B. (2019). The reparameterized unified model system: A diagnostic assessment modeling approach. In *Handbook of Diagnostic Classification Models* 47–79. Springer, Berlin.

TATSUOKA, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In *Diagnostic Monitoring of Skill and Knowledge Acquisition* 453–488.

TATSUOKA, C. (2002). Data analytic methods for latent partially ordered classification models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **51** 337–350. MR1920801 https://doi.org/10.1111/1467-9876.00272

TATSUOKA, K. K. (2009). *Cognitive Assessment*: *An Introduction to the Rule Space Method*. Routledge, New York.

TEMPLIN, J. and BRADSHAW, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika* **79** 317–339. MR3255122 https://doi.org/10.1007/s11336-013-9362-0

TEMPLIN, J. L. and HENSON, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **11** 287–305. https://doi.org/10.1037/1082-989X.11.3.287

TEMPLIN, J. L., HENSON, R. A., TEMPLIN, S. E. and ROUSSOS, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Appl. Psychol. Meas.* **32** 559–574. MR2528297 https://doi.org/10.1177/0146621607300286

VON DAVIER, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **61** 287–307. MR2649038 https://doi.org/10.1348/000711007X193957

XU, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Statist.* **45** 675–707. MR3650397 https://doi.org/10.1214/16-AOS1464

XU, G. and SHANG, Z. (2018). Identifying latent structures in restricted latent class models. *J. Amer. Statist. Assoc.* **113** 1284–1295. MR3862357 https://doi.org/10.1080/01621459.2017.1340889

XU, G. and ZHANG, S. (2016). Identifiability of diagnostic classification models. *Psychometrika* **81** 625–649. MR3535051 https://doi.org/10.1007/s11336-015-9471-z

ZHAO, S., ENGELHARDT, B. E., MUKHERJEE, S. and DUNSON, D. B. (2018). Fast moment estimation for generalized latent Dirichlet models. *J. Amer. Statist. Assoc.* **113** 1528–1540. MR3902227 https://doi.org/10.1080/01621459.2017.1341839

The Annals of Applied Statistics
2021, Vol. 15, No. 3, 1556–1581
https://doi.org/10.1214/21-AOAS1453

ASSESSING SELECTION BIAS IN REGRESSION COEFFICIENTS ESTIMATED FROM NONPROBABILITY SAMPLES WITH APPLICATIONS TO GENETICS AND DEMOGRAPHIC SURVEYS

BY BRADY T. WEST[1,*], RODERICK J. LITTLE[2,‡], REBECCA R. ANDRIDGE[3,‖], PHILIP S. BOONSTRA[2,§], ERIN B. WARE[1,†], ANITA PANDIT[2,¶] AND FERNANDA ALVARADO-LEITON[4]

[1]*Survey Research Center, Institute for Social Research, University of Michigan,* [*]*bwest@umich.edu;* [†]*ebakshis@umich.edu*
[2]*Department of Biostatistics, School of Public Health, University of Michigan,* [‡]*rlittle@umich.edu;* [§]*philb@umich.edu;* [¶]*anitapan@umich.edu*
[3]*Division of Biostatistics, College of Public Health, Ohio State University,* [‖]*andridge.1@osu.edu*
[4]*Michigan Program in Survey and Data Science, Institute for Social Research, University of Michigan, mleiton@umich.edu*

Selection bias is a serious potential problem for inference about relationships of scientific interest based on samples without well-defined probability sampling mechanisms. Motivated by the potential for selection bias in: (a) estimated relationships of polygenic scores (PGSs) with phenotypes in genetic studies of volunteers and (b) estimated differences in subgroup means in surveys of smartphone users, we derive novel measures of selection bias for estimates of the coefficients in linear and probit regression models fitted to nonprobability samples, when aggregate-level auxiliary data are available for the selected sample and the target population. The measures arise from normal pattern-mixture models that allow analysts to examine the sensitivity of their inferences to assumptions about nonignorable selection in these samples. We examine the effectiveness of the proposed measures in a simulation study and then use them to quantify the selection bias in: (a) estimated PGS-phenotype relationships in a large study of volunteers recruited via Facebook and (b) estimated subgroup differences in mean past-year employment duration in a nonprobability sample of low-educated smartphone users. We evaluate the performance of the measures in these applications using benchmark estimates from large probability samples.

REFERENCES

ANDRIDGE, R. R. and LITTLE, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Stat.* **27** 153–180.

ANDRIDGE, R. R. and LITTLE, R. J. (2020). Proxy pattern-mixture analysis for a binary variable subject to nonresponse. *J. Off. Stat.* **36** 703–728.

ANDRIDGE, R. R., WEST, B. T., LITTLE, R. J. A., BOONSTRA, P. S. and ALVARADO-LEITON, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1465–1483. MR4022822 https://doi.org/10.1111/rssc.12371

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A. and TOURANGEAU, R. (2013). Summary report of the AAPOR task force on nonprobability sampling. *J. Sur. Stat. Methodol.* **1** 90–143.

BELSKY, D. W. and ISRAEL, S. (2014). Integrating genetics and social science: Genetic risk scores. *Biodemogr. Soc. Biol.* **60** 137–155.

BLUMBERG, S. and LUKE, J. (2018). Wireless substitution: Early release of estimates from the National Health Interview Survey. Available at https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201812.pdf.

BOONSTRA, P. S., ANDRIDGE, R. R., WEST, B. T., LITTLE, R. J. A. and ALVARADO-LEITON, F. (2021). A simulation study of diagnostics for selection bias. *J. Off. Stat.* (in press).

BRICK, J. M. and WILLIAMS, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *Ann. Am. Acad. Polit. Soc. Sci.* **645** 36–59.

Key words and phrases. Linear regression, probit regression, nonprobability samples, selection bias, polygenic scores, National Survey of Family Growth.

CLIFFORD, S., JEWELL, R. M. and WAGGONER, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Res. Polit.* **2** 2053168015622072.

INTERNATIONAL SCHIZOPHRENIA CONSORTIUM (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748–752.

COUPER, M. P., GREMEL, G., AXINN, W. G., GUYER, H., WAGNER, J. and WEST, B. T. (2018). New options for national population surveys: The implications of Internet and smartphone coverage. *Soc. Sci. Res.* **73** 221–235.

DE LEEUW, E., HOX, J. and LUITEN, A. (2018). International nonresponse trends across countries and years: An analysis of 36 years of Labour Force Survey data. Survey Insights: Methods from the Field. Retrieved from https://surveyinsights.org/?p=10452.

DUDBRIDGE, F. (2016). Polygenic epidemiology. *Genet. Epidemiol.* **40** 268–272.

ELLIOTT, M. R. and VALLIANT, R. (2017). Inference for nonprobability samples. *Statist. Sci.* **32** 249–264. MR3648958 https://doi.org/10.1214/16-STS598

GLYNN, R. J., LAIRD, N. M. and RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.) 115–142. Springer, New York.

GOLDBERGER, A. S. (1981). Linear regression after selection. *J. Econometrics* **15** 357–366. MR0613755 https://doi.org/10.1016/0304-4076(81)90100-7

HAN, J. W., ZHENG, H. F., CUI, Y., SUN, L. D., YE, D. Q., HU, Z. and ZHANG, X. J. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41** 1234–1239.

HECKMAN, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement* **5** 475–492. NBER.

HOULSTON, R. S., CHEADLE, J., DOBBINS, S. E., TENESA, A., JONES, A. M., HOWARTH, K. and TOMLINSON, I. P. M. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42** 973–979.

KAPOOR, M., CHOU, Y. L., EDENBERG, H. J., FOROUD, T., MARTIN, N. G., MADDEN, P. A. F. and AGRAWAL, A. (2016). Genome-wide polygenic scores for age at onset of alcohol dependence and association with alcohol-related measures. *Transl. Psychiatry* **6** e761.

KHOURY, M. J., JANSSENS, A. C. J. and RANSOHOFF, D. F. (2013). How can polygenic inheritance be used in population screening for common diseases? *Genet. Med.* **15** 437–443.

LEWIS, C. M. and VASSOS, E. (2017). Prospects for using risk scores in polygenic medicine. *Gen. Med.* **9** 96.

LINDGREN, C. M., HEID, I. M., RANDALL, J. C., LAMINA, C., STEINTHORSDOTTIR, V., QI, L. and JACKSON, A. U. (2009). Correction: Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5** e1000508.

LITTLE, R. J. (1985). A note about models for selectivity bias. *Econometrica* **53** 1469–1474.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81** 471–483. MR1311091 https://doi.org/10.1093/biomet/81.3.471

LITTLE, R. J. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data*, 3rd ed. Wiley, New York.

LITTLE, R. J. A., WEST, B. T., BOONSTRA, P. and HU, J. (2020). Measures of the degree of departure from ignorable sample selection. *J. Sur. Stat. Methodol.* **8** 932–964.

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E., PERS, T. H., DAY, F. R. and SPELIOTES, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206.

MAHER, B. S. (2015). Polygenic scores in epidemiology: Risk prediction, etiology, and clinical utility. *Curr. Epidemiol. Rep.* **2** 239–244. https://doi.org/10.1007/s40471-015-0055-3

MAITY, A. K., PRADHAN, V. and DAS, U. (2019). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *Amer. Statist.* **73** 340–349. MR4027874 https://doi.org/10.1080/00031305.2017.1407359

MANDEL, H. and SEMYONOV, M. (2014). Gender pay gap and employment sector: Sources of earnings disparities in the United States, 1970-2010. *Demography* **51** 1597–1618.

MARTIN, A. R., KANAI, M., KAMATANI, Y., OKADA, Y., NEALE, B. M. and DALY, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51** 584–591.

MORGAN, J. and DAVID, M. (1963). Education and income. *Q. J. Econ.* **77** 423–437.

MULLER, A. (2002). Education, income inequality, and mortality: A multiple regression analysis. *Br. Med. J.* **324** 23–25.

NALLS, M. A., PANKRATZ, N., LILL, C. M., DO, C. B., HERNANDEZ, D. G., SAAD, M. and SINGLETON, A. B. (2014). Large scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46** 989–993.

NEALE, B. M., MEDLAND, S. E., RIPKE, S., ASHERSON, P., FRANKE, B., LESCH, K. P. and DALY, M. (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J. Am. Acad. Child Adolesc. Psych.* **49** 884–897.

NISHIMURA, R., WAGNER, J. and ELLIOTT, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *Int. Stat. Rev.* **84** 43–62. MR3491278 https://doi.org/10.1111/insr.12100

SCHIZOPHRENIA WORKING GROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511** 421–427.

OKBAY, A., BEAUCHAMP, J. P., FONTANA, M. A., LEE, J. J., PERS, T. H., RIETVELD, C. A. and OSKARSSON, S. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533** 539.

PETROLIA, D. R. and BHATTACHARJEE, S. (2009). Revisiting incentive effects: Evidence from a random-sample mail survey on consumer preferences for fuel ethanol. *Public Opin. Q.* **73** 537–550.

PRESSER, S. and MCCULLOCH, S. (2011). The growth of survey research in the United States: Government-sponsored surveys, 1984–2004. *Soc. Sci. Res.* **40** 1019–1024.

REVILLA, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *Methods Data Anal.* **11** 28.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196 https://doi.org/10.1093/biomet/63.3.581

RYU, E., COUPER, M. P. and MARANS, R. W. (2005). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. *Int. J. Public Opin. Res.* **18** 89–106.

SCHOUTEN, B., COBBEN, F. and BETHLEHEM, J. (2009). Indicators for the representativeness of survey response. *Surv. Methodol.* **35** 101–113.

SKLAR, P., RIPKE, S., SCOTT, L. J., ANDREASSEN, O. A., CICHON, S., CRADDOCK, N. and CORVIN, A. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43** 977–983.

STEIN, M. B., WARE, E. B., MITCHELL, C., CHEN, C. Y., BORJA, S., CAI, T. and JAIN, S. (2017). Genome-wide association studies of suicide attempts in US soldiers. *Am. J. Med. Genet., Part B Neuropsychiatr. Genet.* **174** 786–797.

TORKAMANI, A., WINEINGER, N. E. and TOPOL, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19** 581–590.

VALLIANT, R. (2019). Comparing alternatives for estimation from nonprobability samples. *J. Sur. Stat. Methodol.* https://doi.org/10.1093/jssam/smz003

WARE, E. B., SCHMITZ, L. L., FAUL, J. D., GARD, A., MITCHELL, C., SMITH, J. A. and KARDIA, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *bioRxiv*. Available at https://www.biorxiv.org/content/early/2017/02/05/106062.

WEST, B. T., LITTLE, R. J. A., ANDRIDGE, R. R., BOONSTRA, P. S., WARE, E. B., PANDIT, A. and ALVARADO-LEITON, F. (2021). Supplement to "Assessing Selection Bias in Regression Coefficients Estimated from Nonprobability Samples with Applications to Genetics and Demographic Surveys." https://doi.org/10.1214/21-AOAS1453SUPPA, https://doi.org/10.1214/21-AOAS1453SUPPB

WILLIAMS, D. and BRICK, J. M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *J. Sur. Stat. Methodol.* **6** 186–211.

WRAY, N. R., GODDARD, M. E. and VISSCHER, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17** 1520–1528.

WRAY, N. R., YANG, J., HAYES, B. J., PRICE, A. L., GODDARD, M. E. and VISSCHER, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14** 507–515.

WRAY, N. R., LEE, S. H., MEHTA, D., VINKHUYZEN, A. A., DUDBRIDGE, F. and MIDDELDORP, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55** 1068–1087.

The Annals of Applied Statistics

Next Issues

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Continued

The Annals of Applied Statistics

Next Issues—*Continued*

Continued

The Annals of Applied Statistics

Next Issues—*Continued*