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*

Corrigendum

THE ANNALS OF APPLIED STATISTICS Vol. 17, No. 3, pp. 1801–2700 September 2023

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: Michael Kosorok, Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, Chapel Hill, NC 27599, USA

President-Elect: Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA

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

Executive Secretary: Peter Hoff, Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA

Treasurer: Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

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

IMS PUBLICATIONS

The Annals of Statistics. *Editors*: Enno Mammen, Institute for Mathematics, Heidelberg University, 69120 Heidelberg, Germany. Lan Wang, Miami Herbert Business School, University of Miami, Coral Gables, FL 33124, USA

The Annals of Applied Statistics. *Editor-In-Chief*: Ji Zhu, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, 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*: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Qi-Man Shao, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, P.R. China

Statistical Science. *Editor*: Moulinath Banerjee, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

The IMS Bulletin. *Editor*: Tati Howell, bulletin@imstat.org

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 17, Number 3, September 2023. 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
2023, Vol. 17, No. 3, 1801–1819
https://doi.org/10.1214/22-AOAS1671

REAL-TIME MECHANISTIC BAYESIAN FORECASTS OF COVID-19 MORTALITY

BY GRAHAM C. GIBSON[1,a], NICHOLAS G. REICH[1,b] AND DANIEL SHELDON[2,c]

[1]*School of Public Health and Health Sciences, University of Massachusetts Amherst,* [a]*gcgibson@lanl.gov,* [b]*nick@umass.edu*
[2]*College of Information and Computer Sciences, University of Massachusetts Amherst,* [c]*sheldon@cs.umass.edu*

The COVID-19 pandemic emerged in late December 2019. In the first six months of the global outbreak, the U.S. reported more cases and deaths than any other country in the world. Effective modeling of the course of the pandemic can help assist with public health resource planning, intervention efforts, and vaccine clinical trials. However, building applied forecasting models presents unique challenges during a pandemic. First, case data available to models in real time represent a nonstationary fraction of the true case incidence due to changes in available diagnostic tests and test-seeking behavior. Second, interventions varied across time and geography leading to large changes in transmissibility over the course of the pandemic. We propose a mechanistic Bayesian model that builds upon the classic compartmental susceptible–exposed–infected–recovered (SEIR) model to operationalize COVID-19 forecasting in real time. This framework includes nonparametric modeling of varying transmission rates, nonparametric modeling of case and death discrepancies due to testing and reporting issues, and a joint observation likelihood on new case counts and new deaths; it is implemented in a probabilistic programming language to automate the use of Bayesian reasoning for quantifying uncertainty in probabilistic forecasts. The model has been used to submit forecasts to the U.S. Centers for Disease Control through the COVID-19 Forecast Hub under the name MechBayes. We examine the performance relative to a baseline model as well as alternate models submitted to the forecast hub. Additionally, we include an ablation test of our extensions to the classic SEIR model. We demonstrate a significant gain in both point and probabilistic forecast scoring measures using MechBayes, when compared to a baseline model, and show that MechBayes ranks as one of the top two models out of nine which regularly submitted to the COVID-19 Forecast Hub for the duration of the pandemic, trailing only the COVID-19 Forecast Hub ensemble model of which which MechBayes is a part.

REFERENCES

ABBOTT, S., HELLEWELL, J., THOMPSON, R. N., SHERRATT, K., GIBBS, H. P., BOSSE, N. I., MUNDAY, J. D., MEAKIN, S., DOUGHTY, E. L. et al. (2020). Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res*. **5** 112.

BERTOZZI, A. L., FRANCO, E., MOHLER, G., SHORT, M. B. and SLEDGE, D. (2020). The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci. USA* **117** 16732–16738. MR4242725 https://doi.org/10.1073/pnas.2006520117

BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. Preprint. Available at arXiv:1701.02434.

BOKLER, B. (1993). Chaos and complexity in measles models: A comparative numerical study. *Math. Med. Biol.* **10** 83–95.

BORCHERING, R. K., VIBOUD, C., HOWERTON, E., SMITH, C. P., TRUELOVE, S., RUNGE, M. C., REICH, N. G., CONTAMIN, L., LEVANDER, J. et al. (2021). Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—United States, April–September 2021. *Morb. Mort. Wkly. Rep*. **70** 719.

Key words and phrases. COVID-19 mortality forecasting, compartmental models, Bayesian differential equations.

BRADBURY, J., FROSTIG, R., HAWKINS, P., LEARY, C., MACLAURIN, D., NECULA, G., PASZKE, A., VANDERPLAS, J., WANDERMAN-MILNE, S. and ZHANG, Q. (2018). JAX: Composable transformations of Python+NumPy programs. Available at http://github.com/google/jax.

CAMERON, C. A. and TRIVEDI, P. K. (1986). Econometric models based on count data. Comparisons and applications of some estimators and tests. *J. Appl. Econometrics* **1** 29–53.

CATCHING, A., CAPPONI, S., YEH, M. T., BIANCO, S. and ANDINO, R. (2021). Examining the interplay between face mask usage, asymptomatic transmission, and social distancing on the spread of COVID-19. *Sci. Rep.* **11** 1–11.

CHEN, R. T., RUBANOVA, Y., BETTENCOURT, J. and DUVENAUD, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* 6571–6583.

CRAMER, E. Y., RAY, E. L., LOPEZ, V. K., BRACHER, J., BRENNEN, A., RIVADENEIRA, A. J. C., GERDING, A., GNEITING, T., HOUSE, K. H. et al. (2021a). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. *medRxiv*.

CRAMER, E. Y., HUANG, Y., WANG, Y., RAY, E. L., CORNELL, M., BRACHER, J., BRENNEN, A., RIVADENEIRA, A. J. C., GERDING, A. et al. (2021b). The United States COVID-19 forecast hub dataset. *medRxiv*.

DONG, E., DU, H. and GARDNER, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20** 533–534.

DORMAND, J. R. and PRINCE, P. J. (1980). A family of embedded Runge–Kutta formulae. *J. Comput. Appl. Math.* **6** 19–26. MR0568599 https://doi.org/10.1016/0771-050X(80)90013-3

DUKIC, V., LOPES, H. F. and POLSON, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *J. Amer. Statist. Assoc.* **107** 1410–1426. MR3036404 https://doi.org/10.1080/01621459.2012.713876

FLAXMAN, S., MISHRA, S., GANDY, A., UNWIN, H. J. T., MELLAN, T. A., COUPLAND, H., WHITTAKER, C., ZHU, H., BERAH, T. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** 257–261.

FRASSO, G. and LAMBERT, P. (2016). Bayesian inference in an extended SEIR model with nonparametric disease transmission rate: An application to the Ebola epidemic in Sierra Leone. *Biostatistics* **17** 779–792. MR3604280 https://doi.org/10.1093/biostatistics/kxw027

GIBSON, G. C., REICH, N. G. and SHELDON, D. (2023). Supplement to "Real-time mechanistic Bayesian forecasts of COVID-19 mortality." https://doi.org/10.1214/22-AOAS1671SUPPA, https://doi.org/10.1214/22-AOAS1671SUPPB

GIORDANO, G., BLANCHINI, F., BRUNO, R., COLANERI, P., DI FILIPPO, A., DI MATTEO, A. and COLANERI, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **26** 855–860.

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275 https://doi.org/10.1111/j.1467-9868.2007.00587.x

GRINSZTAJN, L., SEMENOVA, E., MARGOSSIAN, C. C. and RIOU, J. (2021). Bayesian workflow for disease transmission modeling in Stan. *Stat. Med.* **40** 6209–6234. MR4339396 https://doi.org/10.1002/sim.9164

HALL, I., GANI, R., HUGHES, H. and LEACH, S. (2007). Real-time epidemic forecasting for pandemic influenza. *Epidemiol. Infect.* **135** 372–385.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

HOTTA, L. K. (2010). Bayesian melding estimation of a stochastic SEIR model. *Math. Popul. Stud.* **17** 101–111. MR2664909 https://doi.org/10.1080/08898481003689528

JOHNDROW, J., BALL, P., GARGIULO, M. and LUM, K. (2020). Estimating the number of SARS-CoV-2 infections and the impact of mitigation policies in the United States. *Harv. Data Sci. Rev.*.

KARLEN, D. (2020). Characterizing the spread of COVID-19. Preprint. Available at arXiv:2007.07156.

KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci.* **115** 700–721.

KOROLEV, I. (2021). Identification and estimation of the SEIRD epidemic model for COVID-19. *J. Econometrics* **220** 63–85. MR4185125 https://doi.org/10.1016/j.jeconom.2020.07.038

KRANTZ, S. G. and RAO, A. S. S. (2020). Level of under-reporting including under-diagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling. *Infect. Control Hosp. Epidemiol.* **41** 857–859.

LAU, H., KHOSRAWIPOUR, T., KOCBACH, P., ICHII, H., BANIA, J. and KHOSRAWIPOUR, V. (2021). Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology* **27** 110–115. https://doi.org/10.1016/j.pulmoe.2020.05.015

LEKONE, P. E. and FINKENSTÄDT, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* **62** 1170–1177. MR2307442 https://doi.org/10.1111/j.1541-0420.2006.00609.x

LÓPEZ, L. and RODO, X. (2020). A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics. Available at SSRN 3576802.

LUTZ, C. S., HUYNH, M. P., SCHROEDER, M., ANYATONWU, S., DAHLGREN, F. S., DANYLUK, G., FERNANDEZ, D., GREENE, S. K., KIPSHIDZE, N. et al. (2019). Applying infectious disease forecasting to public health: A path forward using influenza forecasting examples. *BMC Public Health* **19** 1659.

MBUVHA, R. and MARWALA, T. (2020). Bayesian inference of COVID-19 spreading rates in South Africa. *PLoS ONE* **15** e0237126. https://doi.org/10.1371/journal.pone.0237126

MIDAS (2020). COVID-19 parameter estimates. Available at https://github.com/midas-network/COVID-19/tree/master/parameter_estimates/2019_novel_coronavirus.

MYERS, M. F., ROGERS, D., COX, J., FLAHAULT, A. and HAY, S. I. (2000). Forecasting disease risk for increased epidemic preparedness in public health. *Adv. Parasitol.* **47** 309–330.

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

ONG, J. B. S., MARK, I., CHEN, C., COOK, A. R., LEE, H. C., LEE, V. J., LIN, R. T. P., TAMBYAH, P. A. and GOH, L. G. (2010). Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS ONE* **5** e10036.

OSTHUS, D., HICKMANN, K. S., CARAGEA, P. C., HIGDON, D. and DEL VALLE, S. Y. (2017). Forecasting seasonal influenza with a state-space SIR model. *Ann. Appl. Stat.* **11** 202–224. MR3634321 https://doi.org/10.1214/16-AOAS1000

PEI, S., KANDULA, S. and SHAMAN, J. (2020). Differential effects of intervention timing on COVID-19 spread in the United States. *medRxiv*. https://doi.org/10.1101/2020.05.15.20103655

PHAN, D., PRADHAN, N. and JANKOWIAK, M. (2019). Composable effects for flexible and accelerated probabilistic programming in NumPyro. Preprint. Available at arXiv:1912.11554.

PREM, K., LIU, Y., RUSSELL, T. W., KUCHARSKI, A. J., EGGO, R. M., DAVIES, N., FLASCHE, S., CLIFFORD, S., PEARSON, C. A. et al. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *Lancet Public Health* **5** E261–E270.

RAHMANDAD, H., LIM, T. Y. and STERMAN, J. (2020). Estimating COVID-19 under-reporting across 86 nations: Implications for projections and control. Available at SSRN 3635047.

RAY, E. L., WATTANACHIT, N., NIEMI, J., KANJI, A. H., HOUSE, K., CRAMER, E. Y., BRACHER, J., ZHENG, A., YAMANA, T. K. et al. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the US. *medRxiv*.

RUSSEL, T. W., HELLEWELL, J., ABBOT, S. et al. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. Available at the Centre for Mathematical Modelling of Infectious Diseases Repository.

RUSSELL, T. W., HELLEWELL, J., JARVIS, C. I., VAN ZANDVOORT, K., ABBOTT, S., RATNAYAKE, R., FLASCHE, S., EGGO, R. M., EDMUNDS, W. J. et al. (2020). Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Euro Surveill.* **25** 2000256.

SHAMAN, J. and KARSPECK, A. (2012). Forecasting seasonal outbreaks of influenza. *Proc. Natl. Acad. Sci. USA* **109** 20425–20430.

SIMONOV, A., SACHER, S. K., DUBÉ, J.-P. H. and BISWAS, S. (2020). The persuasive effect of fox news: Non-compliance with social distancing during the COVID-19 pandemic. Technical report, National Bureau of Economic Research.

SMIRNOVA, A., DECAMP, L. and CHOWELL, G. (2019). Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the SEIR model. *Bull. Math. Biol.* **81** 4343–4365. MR4034830 https://doi.org/10.1007/s11538-017-0284-3

SYAFRUDDIN, S. and NOORANI, M. (2012). SEIR model for transmission of dengue fever in Selangor Malaysia. *Int. J. Mod. Phys. Conf. Ser.* **9** 380–389.

UBER LABS AI (2020). "NumPyro." Available at https://readthedocs.org/projects/numpyro/downloads/pdf/stable/.

WEINBERGER, D. M., CHEN, J., COHEN, T., CRAWFORD, F. W., MOSTASHARI, F., OLSON, D., PITZER, V. E., REICH, N. G., RUSSI, M. et al. (2020). Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. *JAMA Intern. Med.* **180** 1336–1344.

YANG, H. and LEE, J. (2020). Variational Bayes method for ODE parameter estimation with application to time-varying SIR model for COVID-19 epidemic. Preprint. Available at arXiv:2011.09718.

YANG, Z., ZENG, Z., WANG, K., WONG, S.-S., LIANG, W., ZANIN, M., LIU, P., CAO, X., GAO, Z. et al. (2020). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *J. Thorac. Dis.* **12** 165.

ZHUANG, L. and CRESSIE, N. (2014). Bayesian hierarchical statistical SIRS models. *Stat. Methods Appl.* **23** 601–646. MR3278930 https://doi.org/10.1007/s10260-014-0280-9

ZHUANG, L., CRESSIE, N., POMEROY, L. and JANIES, D. (2013). Multi-species SIR models from a dynamical Bayesian perspective. *Theor. Ecol.* **6** 457–473.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1820–1840
https://doi.org/10.1214/22-AOAS1683

CALIBRATION OF SPATIOTEMPORAL FORECASTS FROM CITIZEN SCIENCE URBAN AIR POLLUTION DATA WITH SPARSE RECURRENT NEURAL NETWORKS

BY MATTHEW BONAS[a] AND STEFANO CASTRUCCIO[b]

Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, [a]*mbonas@nd.edu,* [b]*scastruc@nd.edu*

With their continued increase in coverage and quality, data collected from personal air quality monitors has become an increasingly valuable tool to complement existing public health monitoring systems over urban areas. However, the potential of using such "citizen science data" for automatic early warning systems is hampered by the lack of models able to capture the high-resolution, nonlinear spatiotemporal features stemming from local emission sources such as traffic, residential heating and commercial activities. In this work we propose a machine-learning approach to forecast high-frequency spatial fields which has two distinctive advantages from standard neural network methods in time: (1) sparsity of the neural network via a spike-and-slab prior and (2) a small parametric space. The introduction of stochastic neural networks generates additional uncertainty, and in this work we propose a fast approach for ensure that the forecast is correctly assessed (calibration), both marginally and spatially. We focus on assessing exposure to urban air pollution in San Francisco, and our results suggest an improvement of over 58% in the mean squared error over standard time-series approach with a calibrated forecast for up to five days.

REFERENCES

ARAUJO, L. N., BELOTTI, J. T., ALVES, T. A., DE SOUZA TADANO, Y., TROJAN, F. and SIQUEIRA, H. (2020). In Analysis of Regularized Echo State Networks on the Impact of Air Pollutants on Human Health 357–364 Springer.

ARDON-DRYER, K., DRYER, Y., WILLIAMS, J. N. and MOGHIMI, N. (2020). Measurements of PM2.5 with PurpleAir under atmospheric conditions. *Atmos. Meas. Tech.* **13** 5441–5458.

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325 https://doi.org/10.1093/biomet/asr054

BLUNDELL, C., CORNEBISE, J., KAVUKCUOGLU, K. and WIERSTRA, D. (2015). Weight uncertainty in neural network. In *Proceedings of the* 32*nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.). *Proceedings of Machine Learning Research* **37** 1613–1622. PMLR, Lille, France.

BONAS, M. and CASTRUCCIO, S. (2023a). Supplement to "Calibration of SpatioTemporal Forecasts from Citizen Science Urban Air Pollution Data with Sparse Recurrent Neural Networks." https://doi.org/10.1214/22-AOAS1683SUPPA

BONAS, M. and CASTRUCCIO, S. (2023b). R code for "Calibration of spatiotemporal forecasts from citizen science urban air pollution data with sparse recurrent neural networks." https://doi.org/10.1214/22-AOAS1683SUPPB

BRIGGS, D. J., DE HOOGH, C., GULLIVER, J., WILLS, J., ELLIOTT, P., KINGHAM, S. and SMALLBONE, K. (2000). A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci. Total Environ.* **253** 151–167.

BROCKWELL, P. J. and DAVIS, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer, New York.

CAL FIRE (2021). List of wildfires on May 27th. Available at https://www.fire.ca.gov/incidents/2020/5/27/range-fire/. Last accessed 2021/04/16.

CARLSTEN, C., SALVI, S., WONG, G. W. and CHUNG, K. F. (2020). Personal strategies to minimise effects of air pollution on respiratory health: Advice for providers, patients and the public. *Eur. Respir. J.* **55** 1902056.

Key words and phrases. Air pollution, spatiotemporal model, citizen science data, recurrent neural network.

CASTRUCCIO, S., OMBAO, H. and GENTON, M. G. (2018). A scalable multi-resolution spatio-temporal model for brain activation and connectivity in fMRI data. *Biometrics* **74** 823–833. MR3860703 https://doi.org/10.1111/biom.12844

CATLETT, C. E., BECKMAN, P. H., SANKARAN, R. and GALVIN, K. K. (2017). Array of things: A scientific research instrument in the public way: Platform design and early lessons learned. In *SCOPE '*17: *Proceedings of the* 2*nd International Workshop on Science of Smart City Operations and Platforms Engineering* 26–33.

CHO, K., VAN MERRIËNBOER, B., BAHDANAU, D. and BENGIO, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST*-8, *Eighth Workshop on Syntax*, *Semantics and Structure in Statistical Translation* 103–111. Association for Computational Linguistics, Doha, Qatar.

CRESSIE, N. and JOHANNESSON, G. (2008). Fixed rank Kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 209–226. MR2412639 https://doi.org/10.1111/j.1467-9868.2007.00633.x

DE BOOR, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. Springer, New York-Berlin. MR0507062

DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. MR3014996 https://doi.org/10.1093/acprof:oso/9780199641178.001.0001

ENVIROMENTAL PROTECTION AGENCY (2021). NAAQS table. Available at https://www.epa.gov/criteria-air-pollutants/naaqs-table. Last accessed 2021/04/18.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275 https://doi.org/10.1111/j.1467-9868.2007.00587.x

GONON, L. and ORTEGA, J.-P. (2021). Fading memory echo state networks are universal. *Neural Netw.* **138** 10–13.

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

GOODKIND, A. L., TESSUM, C. W., COGGINS, J. S., HILL, J. D. and MARSHALL, J. D. (2019). Fine-scale damage estimates of particulate matter air pollution reveal opportunities for location-specific mitigation of emissions. *Proc. Natl. Acad. Sci. USA* **116** 8775–8780.

GRANGER, C. W. J. and JOYEUX, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Series Anal.* **1** 15–29. MR0605572 https://doi.org/10.1111/j.1467-9892.1980.tb00297.x

GRAVES, A. (2011). Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds.) **24**. Curran Associates, Red Hook.

GRELL, G. A., PECKHAM, S. E., SCHMITZ, R., MCKEEN, S. A., FROST, G., SKAMAROCK, W. C. and EDER, B. (2005). Fully coupled "online" chemistry within the WRF model. *Atmos. Environ.* **39** 6957–6975.

HART, A., HOOK, J. and DAWES, J. (2020). Embedding and approximation theorems for echo state networks. *Neural Netw.* **128** 234–247.

HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. Available at arXiv:1207.0580.

HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Comput.* **9** 1735–1780.

HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176. MR0614953 https://doi.org/10.1093/biomet/68.1.165

HUANG, H., CASTRUCCIO, S. and GENTON, M. G. (2022). Forecasting high-frequency spatio-temporal wind power with dimensionally reduced echo state networks. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 449–466. MR4396917 https://doi.org/10.1111/rssc.12540

HYNDMAN, R. J. and ATHANASOPOULOS, G. (2021). *Forecasting*: *Principles and Practice*, OTexts.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

JAEGER, H. (2001). The "echo state" approach to analysing and training recurrent neural networks-with an erratum note Technology GMD Technical Report 148, German National Research Center for Information, Bonn, Germany.

JAEGER, H. (2007). Echo state network. *Scholarpedia* **2** 2330.

KARIMI, A. and PAUL, M. R. (2010). Extensive chaos in the Lorenz-96 model. *Chaos* **20** 043105.

KELLY, K. E., XING, W. W., SAYAHI, T., MITCHELL, L., BECNEL, T., GAILLARDON, P.-E., MEYER, M. and WHITAKER, R. T. (2021). Community-based measurements reveal unseen differences during air pollution episodes. *Environ. Sci. Technol.* **55** 120–128.

LEVY, R. C., REMER, L. A., KLEIDMAN, R. G., MATTOO, S., ICHOKU, C., KAHN, R. and ECK, T. F. (2010). Global evaluation of the collection 5 modis dark-target aerosol products over land. *Atmos. Chem. Phys.* **10** 10399–10420.

LIU, H., CAI, J., WANG, Y. and ONG, Y. S. (2018). Generalized robust Bayesian committee machine for large-scale Gaussian process regression. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 3131–3140.

LORENZ, E. (1996). Predictability: A problem partly solved. In *Proceedings Seminar on Predictability* 1–18. ECMWF, Reading Berkshire, UK.

LUKOSEVICIUS, M. (2012). A practical guide to applying echo state networks. In *Neural Networks*: *Tricks of the Trade* 659–686. Springer, Berlin.

MALSINER-WALLI, G. and WAGNER, H. (2011). Comparing spike and slab priors for Bayesian variable selection. *Aust. J. Stat.* **40** 241–264.

MCDERMOTT, P. L. and WIKLE, C. K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat* **6** 315–330. MR3716760 https://doi.org/10.1002/sta4.160

MCDERMOTT, P. L. and WIKLE, C. K. (2018). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics* **30** e2553.

MCDERMOTT, P. L. and WIKLE, C. K. (2019). Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy* **21** 184. MR3923929 https://doi.org/10.3390/e21020184

MILLER, J. and SAFFORD, H. (2012). Trends in wildfire severity: 1984 to 2010 in the Sierra Nevada, modoc plateau, and southern cascades, California, USA. *Fire Ecol.* **8** 41–57.

MONTEIRO, A., LOPES, M., MIRANDA, A. I., BORREGO, C. and VAUTARD, R. (2005). Air pollution forecast in Portugal: A demand from the new air quality framework directive. *Int. J. Environ. Pollut.* **25** 1–9.

PACIOREK, C. J. and SCHERVISH, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17** 483–506. MR2240939 https://doi.org/10.1002/env.785

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

RISSER, M. D. and CALDER, C. A. (2017). Local likelihood estimation for covariance functions with spatially-varying parameters: The convospat package for R. *J. Stat. Softw. Artic.* **81** 1–32.

SATISH, L. and GURURAJ, B. (1993). Use of hidden Markov models for partial discharge pattern classification. *IEEE Trans. Electr. Insul.* **28** 172–182.

SHEN, P., CRIPPA, P. and CASTRUCCIO, S. (2021). Assessing urban mortality from wildfires with a citizen science network. *Air Qual. Atmos. Health* **14** 2015–2027.

SORENSEN, A., JORDAN, R., LADEAU, S., BIEHLER, D., WILSON, S., PITAS, J.-H. and LEISNHAM, P. (2019). Reflecting on efforts to design an inclusive citizen science project in West Baltimore. *Citizen Sci. Theory Pract.* **4** 1–13.

SOUTH COAST AIR QUALITY MANAGEMENT DISTRICT (2021). Air quality sensor performance evaluation center. Available at http://www.aqmd.gov/docs/default-source/aq-spec/summary/purpleair-pa-ii---summary-report.pdf?sfvrsn=16. Last accessed 2021/04/14.

VIANNA NETO, J. H., SCHMIDT, A. M. and GUTTORP, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 103–122. MR3148271 https://doi.org/10.1111/rssc.12027

WOLBERG, G. and ALFY, I. (1999). Monotonic cubic spline interpolation. In *Proceedings of Computer Graphics International* 188–195.

WORLD HEALTH ORGANIZATION (2014). *Global Status Report on Noncommunicable Diseases* 2014. World Health Organization, Paris.

XU, X. and REN, W. (2019). Prediction of air pollution concentration based on mrmr and echo state network. *Appl. Sci.* **9** 1811.

ZHANG, Y., WEST, J. J., MATHUR, R., XING, J., HOGREFE, C., ROSELLE, S. J., BASH, J. O., PLEIM, J. E., GAN, C.-M. et al. (2018). Long-term trends in the ambient $pm_{2.5}$- and o_3-related mortality burdens in the United States under emission reductions from 1990 to 2010. *Atmos. Chem. Phys.* **18** 15003–15016.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1841–1860
https://doi.org/10.1214/22-AOAS1686

TRACKING HEMATOPOIETIC STEM CELL EVOLUTION IN A WISKOTT–ALDRICH CLINICAL TRIAL

BY DANILO PELLIN[1,a], LUCA BIASCO[2,b], SERENA SCALA[3,c], CLELIA DI SERIO[4,d] AND ERNST C. WIT[5,e]

[1]*Gene Therapy Program, Harvard Medical School,* [a]*danilo_pellin@dfci.harvard.edu*
[2]*GOS Institute of Child Health, University College London,* [b]*l.biasco@ucl.ac.uk*
[3]*San Raffaele Telethon Institute for Gene Therapy (SR-TIGET), IRCCS San Raffaele Scientific Institute,* [c]*scala.serena@hsr.it*
[4]*University Centre for Statistics in the Biomedical Sciences, Vita-Salute San Raffaele University,* [d]*diserio.clelia@hsr.it*
[5]*Institute of Computing, Università della Svizzera italiana,* [e]*wite@usi.ch*

Hematopoietic stem cells (HSC) are the cells that give rise to all other blood cells and, as such, they are crucial in the healthy development of individuals. Wiskott–Aldrich Syndrome (WAS) is a severe disorder affecting the regulation of hematopoietic cells and is caused by mutations in the WASP gene. We consider data from a revolutionary gene therapy clinical trial, where HSC harvested from three WAS patients' bone marrow have been edited and corrected using viral vectors. Upon reinfusion into the patient, the HSC multiply and differentiate into other cell types. The aim is to unravel the cell multiplication and cell differentiation process, which has until now remained elusive. This paper models the replenishment of blood lineages resulting from corrected HSC via a multivariate, density-dependent Markov process and develops an inferential procedure to estimate the dynamic parameters, given a set of temporally sparsely observed trajectories. Starting from the master equation, we derive a system of nonlinear differential equations for the evolution of the first- and second-order moments over time. We use these moment equations in a generalized method-of-moments framework to perform inference. The performance of our proposal has been evaluated by considering different sampling scenarios and measurement errors of various strengths using a simulation study. We also compared it to another state-of-the-art approach and found that our method is statistically more efficient. By applying our method to the WAS gene therapy data, we found strong evidence for a myeloid-based developmental pathway of hematopoietic cells where fates of lymphoid and myeloid cells remain coupled, even after the loss of erythroid potential. All code used in this manuscript can be found in the online Supplementary Material, and the latest version of the code is available at https://github.com/dp3ll1n/SLCDP_v1.0.

REFERENCES

AIUTI, A., BIASCO, L., SCARAMUZZA, S., FERRUA, F., CICALESE, M. P., BARICORDI, C., DIONISIO, F., CALABRIA, A., GIANNELLI, S. et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott–Aldrich syndrome. *Science* **341** 1233151. https://doi.org/10.1126/science.1233151

BAILEY, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley, New York. MR0165572

BERRY, C. C., GILLET, N. A., MELAMED, A., GORMLEY, N., BANGHAM, C. R. and BUSHMAN, F. D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28** 755–762.

BIASCO, L., AMBROSI, A., PELLIN, D., BARTHOLOMAE, C., BRIGIDA, I., RONCAROLO, M. G., DI SERIO, C., VON KALLE, C., SCHMIDT, M. et al. (2011). Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol. Med.* **2** 1757–4684.

Key words and phrases. Gene therapy, clonal tracking, Wiskott–Aldrich Syndrome, multivariate Markov process, master equation, generalized method-of-moments, nonlinear differential equations.

BIASCO, L., PELLIN, D., SCALA, S., DIONISIO, F., BASSO-RICCI, L., LEONARDELLI, L., SCARAMUZZA, S., BARICORDI, C., FERRUA, F. et al. (2016). In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19** 107–119.

BIFFI, A., MONTINI, E., LORIOLI, L., CESANI, M., FUMAGALLI, F., PLATI, T., BALDOLI, C., MARTINO, S., CALABRIA, A. et al. (2013). Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* **341** 1233158.

BJÖRCK, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA. MR1386889 https://doi.org/10.1137/1.9781611971484

BLANPAIN, C., HORSLEY, V. and FUCHS, E. (2007). Epithelial stem cells: Turning over new leaves. *Cell* **128** 445–458. https://doi.org/10.1016/j.cell.2007.01.014

CALABRIA, A., LEO, S., BENEDICENTI, F., CESANA, D., SPINOZZI, G., ORSINI, M., MERELLA, S., STUPKA, E., ZANETTI, G. et al. (2014). VISPA: A computational pipeline for the identification and analysis of genomic vector integration sites. *Gen. Med.* **6** 1–12.

CATLIN, S. N., BUSQUE, L., GALE, R. E., GUTTORP, P. and ABKOWITZ, J. L. (2011). The replication rate of human hematopoietic stem cells in vivo. *Blood* **117** 4460–4466.

ELF, J. and EHRENBERG, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.* **13** 2475–2484. https://doi.org/10.1101/gr.1196503

GARDINER, C. W. (1985). *Handbook of Stochastic Methods*, 2nd ed. *Springer Series in Synergetics* **13**. Springer, Berlin. MR0858704

GILLESPIE, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81** 2340–2361.

GOLIGHTLY, A. and WILKINSON, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61** 781–788. MR2196166 https://doi.org/10.1111/j.1541-0420.2005.00345.x

GOLIGHTLY, A. and WILKINSON, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Comput. Statist. Data Anal.* **52** 1674–1693. MR2422763 https://doi.org/10.1016/j.csda.2007.05.019

GRIMA, R. (2012). A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.* **136** 04B616.

KAWAMOTO, H., WADA, H. and KATSURA, Y. (2010). A revised scheme for developmental pathways of hematopoietic cells: The myeloid-based model. *Int. Immunol.* **22** 65–70. https://doi.org/10.1093/intimm/dxp125

KEELING, M. J. (2000). Multiplicative moments and measures of persistence in ecology. *J. Theoret. Biol.* **205** 269–281. https://doi.org/10.1006/jtbi.2000.2066

LEIDNER, R., SANJUAN SILVA, N., HUANG, H., SPROTT, D., ZHENG, C., SHIH, Y.-P., LEUNG, A., PAYNE, R., SUTCLIFFE, K. et al. (2022). Neoantigen T-cell receptor gene therapy in pancreatic cancer. *N. Engl. J. Med.* **386** 2112–2119.

LEONARDELLI, L., PELLIN, D., SCALA, S., DIONISIO, F., RICCI, L. B., CITTARO, D., DI SERIO, C., AIUTI, A. and BIASCO, L. (2016). 531. Computational pipeline for the identification of integration sites and novel method for the quantification of clone sizes in clonal tracking studies. *Mol. Ther.* **24** S212–S213.

MEYER-BAHLBURG, A., BECKER-HERMAN, S., HUMBLET-BARON, S., KHIM, S., WEBER, M., BOUMA, G., THRASHER, A. J., BATISTA, F. D. and RAWLINGS, D. J. (2008). Wiskott-Aldrich syndrome protein deficiency in B cells results in impaired peripheral homeostasis. *Blood* **112** 4158–4169.

NÅSELL, I. (2003a). An extension of the moment closure method. *Theor. Popul. Biol.* **64** 233–239.

NÅSELL, I. (2003b). Moment closure and the stochastic logistic model. *Theor. Popul. Biol.* **63** 159–168.

NALDINI, L. (2011). Ex vivo gene transfer and correction for cell-based therapies. *Nat. Rev. Genet.* **12** 301–315. https://doi.org/10.1038/nrg2985

PELLIN, D., BIASCO, L., AIUTI, A., DI SERIO, M. C. and WIT, E. C. (2019). Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking. *Appl. Netw. Sci.* **4** 115.

PELLIN, D., BIASCO, L., SCALA, S., DI SERIO, C. and WIT, E. C. (2023). Supplement to "Tracking hematopoietic stem cell evolution in a Wiskott–Aldrich clinical trial." https://doi.org/10.1214/22-AOAS1686SUPPA, https://doi.org/10.1214/22-AOAS1686SUPPB

RISKEN, H. (1984). *The Fokker–Planck Equation*: *Methods of Solution and Applications*. *Springer Series in Synergetics* **18**. Springer, Berlin. MR0749386 https://doi.org/10.1007/978-3-642-96807-5

SCHNOERR, D., SANGUINETTI, G. and GRIMA, R. (2017). Approximation and inference methods for stochastic biochemical kinetics—A tutorial review. *J. Phys. A* **50** 093001, 60 pp. MR3609073 https://doi.org/10.1088/1751-8121/aa54d9

SENDER, R. and MILO, R. (2021). The distribution of cellular turnover in the human body. *Nat. Med.* **27** 45–48. https://doi.org/10.1038/s41591-020-01182-9

SINGH, A. and HESPANHA, J. P. (2007). A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.* **69** 1909–1925. MR2329186 https://doi.org/10.1007/s11538-007-9198-9

SNIPPERT, H. J. and CLEVERS, H. (2011). Tracking adult stem cells. *EMBO Rep*. **12** 113–122. https://doi.org/10.1038/embor.2010.216

VAN KAMPEN, N. G. (1981). *Stochastic Processes in Physics and Chemistry*. *Lecture Notes in Math*. **888**. North-Holland, Amsterdam. MR0648937

WEISSMAN, I. L. (2000). Stem cells: Units of development, units of regeneration, and units in evolution. *Cell* **100** 157–168. https://doi.org/10.1016/s0092-8674(00)81692-x

WHITTLE, P. (1957). On the use of the normal approximation in the treatment of stochastic processes. *J. Roy. Statist. Soc. Ser. B* **19** 268–281. MR0102131

WILKINSON, D. J. (2006). *Stochastic Modelling for Systems Biology*. *Chapman & Hall/CRC Mathematical and Computational Biology Series*. CRC Press/CRC, Boca Raton, FL. MR2222876

WU, C., LI, B., LU, R., KOELLE, S. J., YANG, Y., JARES, A., KROUSE, A. E., METZGER, M., LIANG, F. et al. (2014). Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14** 486–499.

XU, J., KOELLE, S., GUTTORP, P., WU, C., DUNBAR, C., ABKOWITZ, J. L. and MININ, V. N. (2019). Statistical inference for partially observed branching processes with application to cell lineage tracking of in vivo hematopoiesis. *Ann. Appl. Stat*. **13** 2091–2119. MR4037423 https://doi.org/10.1214/19-aoas1272

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1861–1883
https://doi.org/10.1214/22-AOAS1690

BAYESIAN MODELING OF INTERACTION BETWEEN FEATURES IN SPARSE MULTIVARIATE COUNT DATA WITH APPLICATION TO MICROBIOME STUDY

BY SHUANGJIE ZHANG[1,a], YUNING SHEN[2,c], IRENE A. CHEN[2,d] AND JUHEE LEE[1,b]

[1]*Department of Statistics, University of California Santa Cruz,* [a]*szhan209@ucsc.edu,* [b]*juheelee@soe.ucsc.edu*
[2]*Department of Chemical and Biomolecular Engineering, University of California Los Angeles,* [c]*yshen@chem.ucsb.edu,* [d]*ireneachen@ucla.edu*

Many statistical methods have been developed for the analysis of microbial community profiles, but due to the complexity of typical microbiome measurements, inference of interactions between microbial features remains challenging. We develop a Bayesian zero-inflated rounded log-normal kernel method to model interaction between microbial features in a community using multivariate count data in the presence of covariates and excess zeros. The model carefully constructs the interaction structure by imposing joint sparsity on the covariance matrix of the kernel and obtains a reliable estimate of the structure with a small sample size. The model also includes zero inflation to account for excess zeros observed in data and infers differential abundance of microbial features associated with covariates through log-linear regression. We provide simulation studies and real data analysis examples to demonstrate the developed model. Comparison of the model to a simpler model and popular alternatives in simulation studies shows that, in addition to an added and important insight on the feature interaction, it yields superior parameter estimates and model fit in various settings.

REFERENCES

AGARWAL, D. K., GELFAND, A. E. and CITRON-POUSTY, S. (2002). Zero-inflated models with application to spatial count data. *Environ. Ecol. Stat.* **9** 341–355. MR1951713 https://doi.org/10.1023/A:1020910605990

ALAM, M. T., AMOS, G. C., MURPHY, A. R., MURCH, S., WELLINGTON, E. M. and ARASARADNAM, R. P. (2020). Microbial imbalance in inflammatory bowel disease patients at different taxonomic levels. *Gut Pathogens* **12** 1–8.

ANDRADE, J. C., ALMEIDA, D., DOMINGOS, M., SEABRA, C. L., MACHADO, D., FREITAS, A. C. and GOMES, A. M. (2020). Commensal obligate anaerobic bacteria and health: Production, storage, and delivery strategies. *Front. Bioeng. Biotechnol.* **8** 550. https://doi.org/10.3389/fbioe.2020.00550

BASHAN, A., GIBSON, T. E., FRIEDMAN, J., CAREY, V. J., WEISS, S. T., HOHMANN, E. L. and LIU, Y.-Y. (2016). Universality of human microbial dynamics. *Nature* **534** 259–262.

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

BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048 https://doi.org/10.1080/01621459.2014.960967

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325 https://doi.org/10.1093/biomet/asr054

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973 https://doi.org/10.1198/jasa.2011.tm10155

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281 https://doi.org/10.1007/s00440-014-0562-z

CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172 https://doi.org/10.1214/15-EJS1081

Key words and phrases. Covariance matrix, differential abundance, factor model, joint sparsity, multivariate count data, rounded kernel model, zero inflation.

CAI, Z., ZHU, T., LIU, F., ZHUANG, Z. and ZHAO, L. (2021). Co-pathogens in periodontitis and inflammatory bowel disease. *Frontiers in Medicine* **8**.

CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. MR2896854 https://doi.org/10.1198/jasa.2011.tm10552

CHATTOPADHYAY, S., ARNOLD, J. D., MALAYIL, L., HITTLE, L., MONGODIN, E. F., MARATHE, K. S., GOMEZ-LOBO, V. and SAPKOTA, A. R. (2021). Potential role of the skin and gut microbiota in premenarchal vulvar lichen sclerosus: A pilot case-control study. *PLoS ONE* **16** e0245243. https://doi.org/10.1371/journal.pone.0245243

CONNOR, N., BARBERÁN, A. and CLAUSET, A. (2017). Using null models to infer microbial co-occurrence networks. *PLoS ONE* **12** e0176751. https://doi.org/10.1371/journal.pone.0176751

FANG, H., HUANG, C., ZHAO, H. and DENG, M. (2015). CCLasso: Correlation inference for compositional data through Lasso. *Bioinformatics* **31** 3172–3180.

FAUST, K., SATHIRAPONGSASUTI, J. F., IZARD, J., SEGATA, N., GEVERS, D., RAES, J. and HUTTENHOWER, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8** e1002606.

FRIEDMAN, J. and ALM, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8** e1002687. https://doi.org/10.1371/journal.pcbi.1002687

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

GAO, C. and ZHOU, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** 785–818. MR3325710 https://doi.org/10.1214/14-AOS1268

GRANTHAM, N. S., GUAN, Y., REICH, B. J., BORER, E. T. and GROSS, K. (2020). MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. *J. Amer. Statist. Assoc.* **115** 599–609. MR4107660 https://doi.org/10.1080/01621459.2019.1626242

JIANG, S., XIAO, G., KOH, A. Y., KIM, J., LI, Q. and ZHAN, X. (2021). A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* **22** 522–540. MR4287166 https://doi.org/10.1093/biostatistics/kxz050

JOVEL, J., PATTERSON, J., WANG, W., HOTTE, N., O'KEEFE, S., MITCHEL, T., PERRY, T., KAO, D., MASON, A. L. et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* **7** 459.

KAAKOUSH, N. O. (2015). Insights into the role of erysipelotrichaceae in the human host. *Front. Cell. Infect. Microbiol.* **5** 84. https://doi.org/10.3389/fcimb.2015.00084

KAMNEVA, O. K. (2017). Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS Comput. Biol.* **13** e1005366. https://doi.org/10.1371/journal.pcbi.1005366

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11** e1004226. https://doi.org/10.1371/journal.pcbi.1004226

LEE, J. and SISON-MANGUS, M. (2018). A Bayesian semiparametric regression model for joint analysis of microbiome data. *Front. Microbiol.* **9** 522. https://doi.org/10.3389/fmicb.2018.00522

LI, Q., GUINDANI, M., REICH, B. J., BONDELL, H. D. and VANNUCCI, M. (2017). A Bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Stat. Anal. Data Min.* **10** 393–409. MR3733613 https://doi.org/10.1002/sam.11350

LLOYD-PRICE, J., ARZE, C., ANANTHAKRISHNAN, A. N., SCHIRMER, M., AVILA-PACHECO, J., POON, T. W., ANDREWS, E., AJAMI, N. J., BONHAM, K. S. et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569** 655–662.

LO, C. and MARCULESCU, R. (2018). PGLasso: Microbial Community Detection through Phylogenetic Graphical Lasso. https://arxiv.org/abs/1807.08039v1.

MA, S., REN, B., MALLICK, H., MOON, Y. S., SCHWAGER, E., MAHARJAN, S., TICKLE, T. L., LU, Y., CARMODY, R. N. et al. (2021). A statistical model for describing and simulating microbial community profiles. *PLoS Comput. Biol.* **17** e1008913.

MAO, J., CHEN, Y. and MA, L. (2020). Bayesian graphical compositional regression for microbiome data. *J. Amer. Statist. Assoc.* **115** 610–624. MR4107661 https://doi.org/10.1080/01621459.2019.1647212

MIRSEPASI-LAURIDSEN, H. C., VALLANCE, B. A., KROGFELT, K. A. and PETERSEN, A. M. (2019). *Clin. Microbiol. Rev.* **32**. https://doi.org/10.1128/CMR.00060-18

NITZAN, O., ELIAS, M., CHAZAN, B., RAZ, R. and SALIBA, W. (2013). Clostridium difficile and inflammatory bowel disease: Role in pathogenesis and implications in treatment. *World J. Gastroenterol.* **19** 7577.

PARADA VENEGAS, D. P., LA FUENTE, M. K. D., LANDSKRON, G., GONZÁLEZ, M. J., QUERA, R., DIJKSTRA, G., HARMSEN, H. J. M., FABER, K. N. and HERMOSO, M. A. (2019). Short chain fatty acids (SCFAs)-mediated gut epithelial and immune regulation and its relevance for inflammatory bowel diseases. *Front. Immunol.* **10** 277. https://doi.org/10.3389/fimmu.2019.00277

PARK, J.-U., OH, B., LEE, J. P., CHOI, M.-H., LEE, M.-J. and KIM, B.-S. (2019). Influence of microbiota on diabetic foot wound in comparison with adjacent normal skin based on the clinical features. *BioMed Research International* **2019**.

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

PAULSON, J. N., STINE, O. C., BRAVO, H. C. and POP, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10** 1200–1202.

PROST, V., GAZUT, S. and BRÜLS, T. (2021). A zero inflated log-normal model for inference of sparse microbial association networks. *PLoS Comput. Biol.* **17** e1009089. https://doi.org/10.1371/journal.pcbi.1009089

QIN, J., SHI, X., XU, J., YUAN, S., ZHENG, B., ZHANG, E., HUANG, G., LI, G., JIANG, G. et al. (2021). Characterization of the genitourinary microbiome of 1165 middle-aged and elderly healthy individuals. *Front. Microbiol.* **12**.

REN, B., BACALLADO, S., FAVARO, S., VATANEN, T., HUTTENHOWER, C. and TRIPPA, L. (2017). Bayesian nonparametric mixed effects models in microbiome data analysis. Preprint. Available at arXiv:1711.01241.

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

SCHWAGER, E., MALLICK, H., VENTZ, S. and HUTTENHOWER, C. (2017). A Bayesian method for detecting pairwise associations in compositional data. *PLoS Comput. Biol.* **13** e1005852. https://doi.org/10.1371/journal.pcbi.1005852

SHULER, K., VERBANIC, S., CHEN, I. A. and LEE, J. (2021). A Bayesian nonparametric analysis for zero-inflated multivariate count data with application to microbiome study. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 961–979. MR4318016 https://doi.org/10.1111/rssc.12493

SOKOL, H., SEKSIK, P., FURET, J., FIRMESSE, O., NION-LARMURIER, I., BEAUGERIE, L., COSNES, J., CORTHIER, G., MARTEAU, P. et al. (2009). Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm. Bowel Dis.* **15** 1183–1189.

TANG, Z.-Z. and CHEN, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **20** 698–713. MR4019726 https://doi.org/10.1093/biostatistics/kxy025

VERBANIC, S., SHEN, Y., LEE, J., DEACON, J. M. and CHEN, I. A. (2020). Microbial predictors of healing and short-term effect of debridement on the microbiome of chronic wounds. *NPJ Biofilms Microbiomes* **6** 1–11.

VESTER-ANDERSEN, M., MIRSEPASI-LAURIDSEN, H., PROSBERG, M., MORTENSEN, C., TRÄGER, C., SKOVSEN, K., THORKILGAARD, T., NØJGAARD, C., VIND, I. et al. (2019). Increased abundance of proteobacteria in aggressive Crohn's disease seven years after diagnosis. *Sci. Rep.* **9** 1–10.

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 1–12.

WANG, Z., MAO, J. and MA, L. (2021). Logistic-tree normal model for microbiome compositions. Preprint. Available at arXiv:2106.15051.

WANG, T. and ZHAO, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73** 792–801. MR3713113 https://doi.org/10.1111/biom.12654

WRZOSEK, L., MIQUEL, S., NOORDINE, M.-L., BOUET, S., CHEVALIER-CURT, M. J., ROBERT, V., PHILIPPE, C., BRIDONNEAU, C., CHERBUY, C. et al. (2013). Bacteroides thetaiotaomicron and Faecalibacterium prausnitzii influence the production of mucus glycans and the development of goblet cells in the colonic epithelium of a gnotobiotic model rodent. *BMC Biol.* **11** 1–13.

XIA, F., CHEN, J., FUNG, W. K. and LI, H. (2013). A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics* **69** 1053–1063. MR3146800 https://doi.org/10.1111/biom.12079

XIAOMING, W., JING, L., YUCHEN, P., HUILI, L., MIAO, Z. and JING, S. (2021). Characteristics of the vaginal microbiomes in prepubertal girls with and without vulvovaginitis. *Eur. J. Clin. Microbiol. Infect. Dis.* **40** 1253–1261.

XIE, F., XU, Y., PRIEBE, C. E. and CAPE, J. (2018). Bayesian estimation of sparse spiked covariance matrices in high dimensions. Preprint. Available at arXiv:1808.07433.

ZHANG, X., MALLICK, H., TANG, Z., ZHANG, L., CUI, X., BENSON, A. K. and YI, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinform.* **18** 1–10.

ZHANG, S., SHEN, Y., CHEN, I. A. and LEE, J. (2023). Supplement to "Bayesian modeling of interaction between features in sparse multivariate count data with application to microbiome study." https://doi.org/10.1214/22-AOAS1690SUPPA, https://doi.org/10.1214/22-AOAS1690SUPPB

ZHAO, S., GAO, C., MUKHERJEE, S. and ENGELHARDT, B. E. (2016). Bayesian group factor analysis with structured sparsity. *J. Mach. Learn. Res.* **17** Paper No. 196, 47. MR3580349

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1884–1908
https://doi.org/10.1214/22-AOAS1696

PROBABILISTIC LEARNING OF TREATMENT TREES IN CANCER

BY TSUNG-HUNG YAO[1,a], ZHENKE WU[1,b], KARTHIK BHARATH[2,e], JINJU LI[1,c] AND VEERABHADRAN BALADANDAYUTHAPANI[1,d]

[1]*Department of Biostatistics, University of Michigan at Ann Arbor,* [a]*yaots@umich.edu,* [b]*zhenkewu@umich.edu,* [c]*lijinju@umich.edu,* [d]*veerab@umich.edu*

[2]*School of Mathematical Sciences, University of Nottingham,* [e]*Karthik.Bharath@nottingham.ac.uk*

Accurate identification of synergistic treatment combinations and their underlying biological mechanisms is critical across many disease domains, especially cancer. In translational oncology research, preclinical systems, such as patient-derived xenografts (PDX), have emerged as a unique study design evaluating multiple treatments administered to samples from the same human tumor implanted into genetically identical mice. In this paper we propose a novel Bayesian probabilistic tree-based framework for PDX data to investigate the hierarchical relationships between treatments by inferring treatment cluster trees, referred to as treatment trees (R_x-tree). The framework motivates a new metric of mechanistic similarity between two or more treatments, accounting for inherent uncertainty in tree estimation; treatments with a high estimated similarity have potentially high mechanistic synergy. Building upon Dirichlet diffusion trees, we derive a closed-form marginal likelihood, encoding the tree structure, which facilitates computationally efficient posterior inference via a new two-stage algorithm. Simulation studies demonstrate superior performance of the proposed method in recovering the tree structure and treatment similarities. Our analyses of a recently collated PDX dataset produce treatment similarity estimates that show a high degree of concordance with known biological mechanisms across treatments in five different cancers. More importantly, we uncover new and potentially effective combination therapies that confer synergistic regulation of specific downstream biological pathways for future clinical investigations. Our accompanying code, data, and shiny application for visualization of results are available at: https://github.com/bayesrx/RxTree.

REFERENCES

ABDOLAHI, S., GHAZVINIAN, Z., MUHAMMADNEJAD, S., SALEH, M., AGHDAEI, H. A. and BAGHAEI, K. (2022). Patient-derived xenograft (PDX) models, applications and challenges in cancer research. *J. Transl. Med.* **20** 206. https://doi.org/10.1186/s12967-022-03405-8

BALKO, J. M., MILLER, T. W., MORRISON, M. M., HUTCHINSON, K., YOUNG, C., RINEHART, C., SÁNCHEZ, V., JEE, D., POLYAK, K. et al. (2012). The receptor tyrosine kinase ErbB3 maintains the balance between luminal and basal breast epithelium. *Proc. Natl. Acad. Sci. USA* **109** 221–226.

BAYAT MOKHTARI, R., HOMAYOUNI, T. S., BALUCH, N., MORGATSKAYA, E., KUMAR, S., DAS, B. and YEGER, H. (2017). Combination therapy in combating cancer. *Oncotarget* **8** 38022–38043.

BERTOTTI, A., MIGLIARDI, G., GALIMI, F., SASSI, F., TORTI, D., ISELLA, C., CORÀ, D., DI NICOLANTONIO, F., BUSCARINO, M. et al. (2011). A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov.* **1** 508–523.

BHIMANI, J., BALL, K. and STEBBING, J. (2020). Patient-derived xenograft models-the future of personalised cancer treatment. *Br. J. Cancer* **122** 601–602. https://doi.org/10.1038/s41416-019-0678-0

BONELLI, M. A., DIGIACOMO, G., FUMAROLA, C., ALFIERI, R., QUAINI, F., FALCO, A., MADEDDU, D., LA MONICA, S., CRETELLA, D. et al. (2017). Combined inhibition of CDK4/6 and PI3K/AKT/mTOR pathways induces a synergistic anti-tumor effect in malignant pleural mesothelioma cells. *Neoplasia* **19** 637–648.

Key words and phrases. Approximate Bayesian computation, Dirichlet diffusion trees, patient derived xenograft, precision medicine, tree-based clustering.

BRAVO, H. C., WRIGHT, S., ENG, K. H., KELES, S. and WAHBA, G. (2009). Estimating tree-structured covariance matrices via mixed-integer programming. *J. Mach. Learn. Res.* **5** 41–48.

CARDONA, G., MIR, A., ROSSELLÓ, F., ROTGER, L. and SÁNCHEZ, D. (2013). Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinform.* **14** 3. https://doi.org/10.1186/1471-2105-14-3

CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference. The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. MR1051420

CLOHESSY, J. G. and PANDOLFI, P. P. (2015). Mouse hospital and co-clinical trial project–from bench to bedside. *Nat. Rev. Clin. Oncol.* **12** 491–498. https://doi.org/10.1038/nrclinonc.2015.62

DAGOGO-JACK, I. and SHAW, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15** 81–94. https://doi.org/10.1038/nrclinonc.2017.166

DOBROLECKI, L. E., AIRHART, S. D., ALFEREZ, D. G., APARICIO, S., BEHBOD, F., BENTIRES-ALJ, M., BRISKEN, C., BULT, C. J., CAI, S. et al. (2016). Patient-derived xenograft (PDX) models in basic and translational breast cancer research. *Cancer Metastasis Rev.* **35** 547–573.

DUMMER, R., ASCIERTO, P. A., GOGAS, H. J., ARANCE, A., MANDALA, M., LISZKAY, G., GARBE, C., SCHADENDORF, D., KRAJSOVA, I. et al. (2018a). Encorafenib plus binimetinib versus vemurafenib or encorafenib in patients with BRAF-mutant melanoma (COLUMBUS): A multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **19** 603–615.

DUMMER, R., ASCIERTO, P. A., GOGAS, H. J., ARANCE, A., MANDALA, M., LISZKAY, G., GARBE, C., SCHADENDORF, D., KRAJSOVA, I. et al. (2018b). Overall survival in patients with BRAF-mutant melanoma receiving encorafenib plus binimetinib versus vemurafenib or encorafenib (COLUMBUS): A multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol.* **19** 1315–1327.

FERLAY, J., ERVIK, M., LAM, F., COLOMBET, M., MERY, L., PIÑEROS, M., ZNAOR, A., SOERJOMATARAM, I. and BRAY, F. (2020). *Global Cancer Observatory*: *Cancer Today*. International Agency for Research on Cancer, Lyon, France. Available from: https://gco.iarc.fr/today, accessed 05.28.2021.

GAO, H., KORN, J. M., FERRETTI, S., MONAHAN, J. E., WANG, Y., SINGH, M., ZHANG, C., SCHNELL, C., YANG, G. et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21** 1318–1325.

GEEL, R. V., ELEZ, E., BENDELL, J. C., FARIS, J. E., LOLKEMA, M. P. J. K., ESKENS, F., SPREAFICO, A., KAVAN, P., DELORD, J.-P. et al. (2014). Phase I study of the selective BRAFV600 inhibitor encorafenib (LGX818) combined with cetuximab and with or without the α-specific PI3K inhibitor BYL719 in patients with advanced BRAF-mutant colorectal cancer. *J. Clin. Oncol.* **32** 3514–3514. https://doi.org/10.1200/jco.2014.32.15\protect\T1\textunderscoresuppl.3514

GRANT, R. L., COMBS, A. B. and ACOSTA, D. (2010). Experimental models for the investigation of toxicological mechanisms. In *Comprehensive Toxicology*, 2nd ed. (C. A. McQueen, ed.) 203–224. Elsevier, Oxford. https://doi.org/10.1016/B978-0-08-046884-6.00110-X

GROISBERG, R. and SUBBIAH, V. (2021). Combination therapies for precision oncology: The ultimate whack-a-mole game. *Clin. Cancer Res.* **27** 2672–2674.

HEAUKULANI, C., KNOWLES, D. A. and GHAHRAMANI, Z. (2014). Beta diffusion trees. In *Proceedings of the* 31*st International Conference on International Conference on Machine Learning—Volume* 32. *ICML*'14 II–1809–II–1817. JMLR.org, Bejing, China.

HIDALGO, M., AMANT, F., BIANKIN, A. V., BUDINSKÁ, E., BYRNE, A. T., CALDAS, C., CLARKE, R. B., DE JONG, S., JONKERS, J. et al. (2014). Patient-derived xenograft models: An emerging platform for translational cancer research. *Cancer Discov.* **4** 998–1013.

KNOWLES, D. A., GAEL, J. V. and GHAHRAMANI, Z. (2011). Message passing algorithms for Dirichlet diffusion trees. In *International Conference on Machine Learning* (*ICML*).

KNOWLES, D. A. and GHAHRAMANI, Z. (2015). Pitman Yor diffusion trees for Bayesian hierarchical clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 271–289. https://doi.org/10.1109/TPAMI.2014.2313115

KNUTH, D. E. (1976). Big omicron and big omega and big theta. *ACM SIGACT News* **8** 18–24.

KOGA, Y. and OCHIAI, A. (2019). Systematic review of patient-derived xenograft models for preclinical studies of anti-cancer drugs in solid tumors. *Cells* **8**. https://doi.org/10.3390/cells8050418

KONOPLEVA, M., MARTINELLI, G., DAVER, N., PAPAYANNIDIS, C., WEI, A., HIGGINS, B., OTT, M., MASCARENHAS, J. and ANDREEFF, M. (2020). MDM2 inhibition: An important step forward in cancer therapy. *Leukemia* **34** 2858–2874.

KRUMBACH, R., SCHÜLER, J., HOFMANN, M., GIESEMANN, T., FIEBIG, H. H. and BECKERS, T. (2011). Primary resistance to cetuximab in a panel of patient-derived tumour xenograft models: Activation of MET as one mechanism for drug resistance. *Eur. J. Cancer* **47** 1231–1243.

KURTZEBORN, K., KWON, H. N. and KUURE, S. (2019). MAPK/ERK signaling in regulation of renal differentiation. *Int. J. Mol. Sci.* **20**. https://doi.org/10.3390/ijms20071779

LAI, Y., WEI, X., LIN, S., QIN, L., CHENG, L. and LI, P. (2017). Current status and perspectives of patient-derived xenograft models in cancer research. *J. Hematol. Oncol.* **10** 106.

LAPOINTE, F.-J. and LEGENDRE, P. (1991). The generation of random ultrametric matrices representing dendrograms. *J. Classification* **8** 177–200. https://doi.org/10.1007/BF02616238

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740

MATHAI, A. M. (1980). Moments of the trace of a noncentral Wishart matrix. *Comm. Statist. Theory Methods* **9** 795–801. MR0573114 https://doi.org/10.1080/03610928008827921

MCCULLAGH, P. (2006). Structured covariance matrices in multivariate regression models Technical Report Department of Statistics, Univ. Chicago.

MÉZARD, M. and MONTANARI, A. (2009). *Information, Physics, and Computation. Oxford Graduate Texts*. Oxford Univ. Press, Oxford. MR2518205 https://doi.org/10.1093/acprof:oso/9780198570837.001.0001

NARAYAN, R. S., MOLENAAR, P., TENG, J., CORNELISSEN, F. M. G., ROELOFS, I., MENEZES, R., DIK, R., LAGERWEIJ, T., BROERSMA, Y. et al. (2020). A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nat. Commun.* **11** 2935. https://doi.org/10.1038/s41467-020-16735-2

NEAL, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 619–629. Oxford Univ. Press, New York. MR2003526

NUNES, M., VRIGNAUD, P., VACHER, S., RICHON, S., LIÈVRE, A., CACHEUX, W., WEISWALD, L.-B., MASSONNET, G., CHATEAU-JOUBERT, S. et al. (2015). Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data. *Cancer Res.* **75** 1560–1566. https://doi.org/10.1158/0008-5472.CAN-14-1590

OH, D.-Y. and BANG, Y.-J. (2020). HER2-targeted therapies—a role beyond breast cancer. *Nat. Rev. Clin. Oncol.* **17** 33–48. https://doi.org/10.1038/s41571-019-0268-3

RASHID, N. U., LUCKETT, D. J., CHEN, J., LAWSON, M. T., WANG, L., ZHANG, Y., LABER, E. B., LIU, Y., YEH, J. J. et al. (2020). High-dimensional precision medicine from patient-derived xenografts. *J. Amer. Statist. Assoc.* **0** 1–15. https://doi.org/10.1080/01621459.2020.1828091

REPETTO, M. V., WINTERS, M. J., BUSH, A., REITER, W., HOLLENSTEIN, D. M., AMMERER, G., PRYCIAK, P. M. and COLMAN-LERNER, A. (2018). CDK and MAPK synergistically regulate signaling dynamics via a shared multi-site phosphorylation region on the scaffold protein Ste5. *Mol. Cell* **69** 938–952.

ROBERT, C., GROB, J. J., STROYAKOVSKIY, D., KARASZEWSKA, B., HAUSCHILD, A., LEVCHENKO, E., SILENI, V. C., SCHACHTER, J., GARBE, C. et al. (2019). Five-year outcomes with dabrafenib plus trametinib in metastatic melanoma. *N. Engl. J. Med.* **381** 626–636. https://doi.org/10.1056/NEJMoa1904059

SAWYERS, C. L. (2013). Perspective: Combined forces. *Nature* **498** S7. https://doi.org/10.1038/498S7a

SISSON, S. A., FAN, Y. and BEAUMONT, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton.

SOKAL, R. R. and ROHLF, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon* **11** 33–40.

SUN, W., SANDERSON, P. E. and ZHENG, W. (2016). Drug combination therapy increases successful drug repositioning. *Drug Discov. Today* **21** 1189–1195.

TENTLER, J. J., TAN, A. C., WEEKES, C. D., JIMENO, A., LEONG, S., PITTS, T. M., ARCAROLI, J. J., MESSERSMITH, W. A. and ECKHARDT, S. G. (2012). Patient-derived tumour xenografts as models for oncology drug development. *Nat. Rev. Clin. Oncol.* **9** 338–350. https://doi.org/10.1038/nrclinonc.2012.61

TOPP, M. D., HARTLEY, L., COOK, M., HEONG, V., BOEHM, E., MCSHANE, L., PYMAN, J., MCNALLY, O., ANANDA, S. et al. (2014). Molecular correlates of platinum response in human high-grade serous ovarian cancer patient-derived xenografts. *Mol. Oncol.* **8** 656–668. https://doi.org/10.1016/j.molonc.2014.01.008

TURNER, B. M., SEDERBERG, P. B., BROWN, S. D. and STEYVERS, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychol. Methods* **18** 368–384.

VAN GEEL, R. M. J. M., TABERNERO, J., ELEZ, E., BENDELL, J. C., SPREAFICO, A., SCHULER, M., YOSHINO, T., DELORD, J.-P., YAMADA, Y. et al. (2017). A phase Ib dose-escalation study of encorafenib and cetuximab with or without alpelisib in metastatic. *Cancer Discov.* **7** 610–619. https://doi.org/10.1158/2159-8290.CD-16-0795

VORA, S. R., JURIC, D., KIM, N., MINO-KENUDSON, M., HUYNH, T., COSTA, C., LOCKERMAN, E. L., POLLACK, S. F., LIU, M. et al. (2014). CDK 4/6 inhibitors sensitize PIK3CA mutant breast cancer to PI3K inhibitors. *Cancer Cell* **26** 136–149.

YAO, T.-H, WU, Z., BHARATH, K., LI, J. and BALADANDAYUTHAPANI, V. (2023). Supplement to "Probabilistic learning of treatment trees in cancer." https://doi.org/10.1214/22-AOAS1696SUPPA, https://doi.org/10.1214/22-AOAS1696SUPPB

YOSHIDA, G. J. (2020). Applications of patient-derived tumor xenograft models and tumor organoids. *J. Hematol. Oncol.* **13** 4. https://doi.org/10.1186/s13045-019-0829-z

YUAN, Y., WEN, W., YOST, S. E., XING, Q., YAN, J., HAN, E. S., MORTIMER, J. and YIM, J. H. (2019). Combination therapy with BYL719 and LEE011 is synergistic and causes a greater suppression of p-S6 in triple negative breast cancer. *Sci. Rep.* **9** 7509.

ZHANG, X., CLAERHOUT, S., PRAT, A., DOBROLECKI, L. E., PETROVIC, I., LAI, Q., LANDIS, M. D., WIECHMANN, L., SCHIFF, R. et al. (2013). A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. *Cancer Res.* **73** 4885–4897. https://doi.org/10.1158/0008-5472.CAN-12-4081

ZHAO, Y., SHUEN, T. W. H., TOH, T. B., CHAN, X. Y., LIU, M., TAN, S. Y., FAN, Y., YANG, H., LYER, S. G. et al. (2018). Development of a new patient-derived xenograft humanised mouse model to study human-specific tumour microenvironment and immunotherapy. *Gut* **67** 1845–1854.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1909–1935
https://doi.org/10.1214/22-AOAS1697

SEQUENTIALLY VALID TESTS FOR FORECAST CALIBRATION

BY SEBASTIAN ARNOLD[a], ALEXANDER HENZI[b] AND JOHANNA F. ZIEGEL[c]

Institute of Mathematical Statistics and Actuarial Science, University of Bern, [a]*sebastian.arnold@stat.unibe.ch,* [b]*alexander.henzi@stat.unibe.ch,* [c]*johanna.ziegel@stat.unibe.ch*

Forecasting and forecast evaluation are inherently sequential tasks. Predictions are often issued on a regular basis, such as every hour, day, or month, and their quality is monitored continuously. However, the classical statistical tools for forecast evaluation are static, in the sense that statistical tests for forecast calibration are only valid if the evaluation period is fixed in advance. Recently, e-values have been introduced as a new, dynamic method for assessing statistical significance. An e-value is a nonnegative random variable with expected value, at most, one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a conservative p-value. Since they naturally lead to statistical tests which are valid under optional stopping, e-values are particularly suitable for sequential forecast evaluation. This article proposes e-values for testing probabilistic calibration of forecasts which is one of the most important notions of calibration. The proposed methods are also more generally applicable for sequential goodness-of-fit testing. We demonstrate in a simulation study that the e-values are competitive in terms of power, when compared to extant methods which do not allow for sequential testing. In this context we introduce test power heat matrices, a graphical tool to compactly visualize results of simulation studies on test power. In a case study we show that the e-values provide important and new useful insights in the evaluation of probabilistic weather forecasts.

REFERENCES

ANDERSON, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9** 1518–1530.

ARNOLD, S., HENZI, A. and ZIEGEL, J. F. (2023). Supplement to "Sequentially valid tests for forecast calibration." https://doi.org/10.1214/22-AOAS1697SUPPA, https://doi.org/10.1214/22-AOAS1697SUPPB, https://doi.org/10.1214/22-AOAS1697SUPPC

BAUER, P., THORPE, A. and BRUNET, G. (2015). The quiet revolution of numerical weather prediction. *Nature* **525** 47–55. https://doi.org/10.1038/nature14956

BOUGEAULT, P., TOTH, Z., BISHOP, C., BROWN, B., BURRIDGE, D., CHEN, D. H., EBERT, B., FUENTES, M., HAMILL, T. M. et al. (2010). The THORPEX interactive grand global ensemble. *Bull. Am. Meteorol. Soc.* **91** 1059–1072.

BUIZZA, R., HOUTEKAMER, P. L., PELLERIN, G., TOTH, Z., ZHU, Y. and WEI, M. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133** 1076–1097.

CHOE, Y. J. and RAMDAS, A. (2021). Comparing sequential forecasters. Preprint, arXiv:2110.00115.

DAWID, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

DIEBOLD, F. X., GUNTHER, T. A. and TAY, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Internat. Econom. Rev.* **39** 863–883.

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 243–268. MR2325275 https://doi.org/10.1111/j.1467-9868.2007.00587.x

GNEITING, T. and RANJAN, R. (2013). Combining predictive distributions. *Electron. J. Stat.* **7** 1747–1782. MR3080409 https://doi.org/10.1214/13-EJS823

Key words and phrases. Forecast calibration, probabilistic calibration, rank histogram, e-value, sequential inference.

GNEITING, T. and RESIN, J. (2021). Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. Preprint, arXiv:2108.03210.
GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957). MR0093415 https://doi.org/10.1080/03461238.1956.10414944
GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2019). Safe testing. Preprint, arXiv:1906.07801.
HAMILL, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129** 550–560.
HEMRI, S., SCHEUERER, M., PAPPENBERGER, F., BOGNER, K. and HAIDEN, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.* **41** 9197–9205.
HENZI, A., MOESCHING, A. and DUEMBGEN, L. (2020). Accelerating the pool-adjacent-violators algorithm for isotonic distributional regression. Preprint, arXiv:2006.05527.
HENZI, A. and ZIEGEL, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika* **109** 647–663. MR4472840 https://doi.org/10.1093/biomet/asab047
HOWARD, S. R. and RAMDAS, A. (2022). Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli* **28** 1704–1728. MR4411508 https://doi.org/10.3150/21-bej1388
JONES, M. C. and FOSTER, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica* **6** 1005–1013. MR1422417
KELLY, J. L. JR. (1956). A new interpretation of information rate. *Bell Syst. Tech. J.* **35** 917–926. MR0090494 https://doi.org/10.1002/j.1538-7305.1956.tb03809.x
MESSNER, J. W., MAYR, G. J. and ZEILEIS, A. (2016). Heteroscedastic censored and truncated regression with crch. *R J.* **8** 173–181.
MOLTENI, F., BUIZZA, R., PALMER, T. N. and PETROLIAGIS, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122** 73–119.
MÜLLER, H.-G. and WANG, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50** 61–76. MR1279435 https://doi.org/10.2307/2533197
PAPADAKIS, M., TSAGRIS, M., DIMITRIADIS, M., FAFALIOS, S., TSAMARDINOS, I., FASIOLO, M., BORBOUDAKIS, G., BURKARDT, J., ZOU, C. et al. (2020). Rfast: A collection of efficient and extremely fast R functions. R package version 2.0.1.
RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. Preprint, arXiv:2009.03167.
RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *Internat. J. Approx. Reason.* **141** 83–109. MR4364897 https://doi.org/10.1016/j.ijar.2021.06.017
SANTAFE, G., CALVO, B., PEREZ, A. and LOZANO, J. A. (2015). bde: Bounded density estimation. R package version 1.0.1.
SHAFER, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *J. Roy. Statist. Soc. Ser. A* **184** 407–478. MR4255905 https://doi.org/10.1111/rssa.12647
SHAFER, G., SHEN, A., VERESHCHAGIN, N. and VOVK, V. (2011). Test martingales, Bayes factors and p-values. *Statist. Sci.* **26** 84–101. MR2849911 https://doi.org/10.1214/10-STS347
STELLATO, B., BANJAC, G., GOULART, P. and BOYD, S. (2019). osqp: Quadratic programming solver using the 'OSQP' library. R package version 0.6.0.3.
STRÄHL, C. and ZIEGEL, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Stat.* **11** 608–639. MR3619318 https://doi.org/10.1214/17-EJS1244
SWINBANK, R., KYOUDA, M., BUCHANAN, P., FROUDE, L., HAMILL, T. M., HEWSON, T. D., KELLER, J. H., MATSUEDA, M., METHVEN, J. et al. (2016). The TIGGE project and its achievements. *Bull. Am. Meteorol. Soc.* **97** 49–67.
THORARINSDOTTIR, T. L., SCHEUERER, M. and HEINZ, C. (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Statist.* **25** 105–122. MR3474039 https://doi.org/10.1080/10618600.2014.977447
TSYPLAKOV, A. (2011). Evaluating density forecasts: A comment. SSRN 1907799.
TURNBULL, B. C. and GHOSH, S. K. (2014). Unimodal density estimation using Bernstein polynomials. *Comput. Statist. Data Anal.* **72** 13–29. MR3139345 https://doi.org/10.1016/j.csda.2013.10.021
VANNITSEM, S., WILKS, D. S. and MESSNER, J., eds. (2018). *Statistical Postprocessing of Ensemble Forecasts* Elsevier, Amsterdam.
VOGEL, P., KNIPPERTZ, P., FINK, A. H., SCHLUETER, A. and GNEITING, T. (2018). Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather Forecast.* **33** 369–388.
VOVK, V. (2021). Testing randomness online. *Statist. Sci.* **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817
VOVK, V. and WANG, R. (2020). Combining p-values via averaging. *Biometrika* **107** 791–808. MR4186488 https://doi.org/10.1093/biomet/asaa027

VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *Ann. Statist.* **49** 1736–1754. MR4298879 https://doi.org/10.1214/20-aos2020

WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

WALD, A. (1947). *Sequential Analysis*. Wiley, New York; CRC Press, London. MR0020764

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. *Monographs on Statistics and Applied Probability* **60**. CRC Press, London. MR1319818 https://doi.org/10.1007/978-1-4899-4493-1

WAND, M. P. and JONES, M. C. (2021). KernSmooth: Functions for kernel smoothing supporting Wand & Jones (1995). R package version 2.23-20.

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA* **117** 16880–16890. MR4242731 https://doi.org/10.1073/pnas.1922664117

WAUDBY-SMITH, I. and RAMDAS, A. (2020). Estimating means of bounded random variables by betting. Preprint, arXiv:2010.09686.

ZIEGEL, J. F. and GNEITING, T. (2014). Copula calibration. *Electron. J. Stat.* **8** 2619–2638. MR3291527 https://doi.org/10.1214/14-EJS964

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1936–1957
https://doi.org/10.1214/22-AOAS1698

BAYESIAN ADDITIVE REGRESSION TREES FOR GENOTYPE BY ENVIRONMENT INTERACTION MODELS

BY DANILO A. SARTI[1,a], ESTEVÃO B. PRADO[1,2,b], ALAN N. INGLIS[1,2,c], ANTÔNIA A. L. DOS SANTOS[1,d], CATHERINE B. HURLEY[1,e], RAFAEL A. MORAL[1,f] AND ANDREW C. PARNELL[1,2,g]

[1]*Insight Centre for Data Analytics, Maynooth University*

[2]*Hamilton Institute, Department of Mathematics and Statistics, Maynooth University,* [a]*daniloasarti@gmail.com,* [b]*estevao.prado@hotmail.com,* [c]*alan.inglis@mu.ie,* [d]*antonia.lemosdossantos.2020@mumail.ie,* [e]*catherine.hurley@mu.ie,* [f]*rafael.deandrademoral@mu.ie,* [g]*Andrew.parnell@mu.ie*

We propose a new class of models for the estimation of genotype by environment (GxE) interactions in plant-based genetics. Our approach, named AMBARTI, uses semiparametric Bayesian additive regression trees to accurately capture marginal genotypic and environment effects along with their interaction in a cut Bayesian framework. We demonstrate that our approach is competitive or superior to similar models widely used in the literature via both simulation and a real world dataset. Furthermore, we introduce new types of visualisation to properly assess both the marginal and interactive predictions from the model. An R package that implements our approach is also available at `https://github.com/ebprado/ambarti`.

REFERENCES

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

ALLARD, R. W. and BRADSHAW, A. D. (1992). Implications of genotype environmental interactions in applied plant breeding. *Crop Sci.* **4** 503–508.

ANBESSA, Y., JUSKIW, P., GOOD, A., NYACHIRO, J. and HELM, J. (2009). Genetic variability in nitrogen use efficiency of spring barley. *Crop Sci.* **49** 1259–1269.

BADU-APRAKU, B., OYEKUNLE, M., OBENG-ANTWI, K., OSUMAN, A., ADO, S., COULIBAY, N., YALLOU, C., ABDULAI, M., BOAKYEWAA, G. et al. (2012). Performance of extra-early maize cultivars based on GGE biplot and AMMI analysis. *J. Agric. Sci.* **150** 473.

BASAK, P., LINERO, A., SINHA, D. and LIPSITZ, S. (2022). Semiparametric analysis of clustered interval-censored survival data using soft Bayesian additive regression trees (SBART). *Biometrics* **78** 880–893. MR4493495 https://doi.org/10.1111/biom.13478

BASFORD, K., KROONENBERG, P. and DELACY, I. (1991). Three-way methods for multiattribute genotype × environment data: An illustrated partial survey. *Field Crops Res.* **27** 131–157.

BRANCOURT-HULMEL, M. and LECOMTE, C. (2003). Effect of environmental variates on genotype × environment interaction of winter wheat: A comparison of biadditive factorial regression to AMMI. *Crop Sci.* **43** 608–617.

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

CROSSA, J., PEREZ-ELIZALDE, S., JARQUIN, D., COTES, J. M., VIELE, K., LIU, G. and CORNELIUS, P. L. (2011). Bayesian estimation of the additive main effects and multiplicative interaction model. *Crop Sci.* **51** 1458–1469.

DE MENDIBURU, F. (2019). Package 'agricolae'. *R Package*, *Version* 1–2.

DENISON, D. G., MALLICK, B. K. and SMITH, A. F. (1998). Bayesian Mars. *Stat. Comput.* **8** 337–346.

DIAS, C. (2005). Métodos para escolha de componentes em modelo de efeito principal aditivo e interação multiplicativa (AMMI). 2005. 73p Ph.D. thesis Tese (Livre Docência)–Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

Key words and phrases. Bayesian nonparametric regression, Bayesian additive regression trees, additive main effects multiplicative interactions model, genotype-by-environment interactions.

DIAS, C. T. D. S. and KRZANOWSKI, W. J. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Sci. Agric.* **63** 169–175.
DORIE, V. (2020). dbarts: Discrete Bayesian Additive Regression Trees Sampler. R package version 0.9-19.
FALCONER, D. and MACKAY, T. (1996). *Introduction to Quantitative Genetics*. Longmans Green, Harlow, Essex, UK.
FARSHADFAR, E. and SUTKA, J. (2003). Locating QTLs controlling adaptation in wheat using AMMI model. *Cereal Res. Commun.* **31** 249–256.
FRANCOM, D. and SANSÓ, B. (2020). BASS: An R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces. *J. Stat. Softw.* **94** 1–36.
GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467. MR0312645 https://doi.org/10.1093/biomet/58.3.453
GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic Simulation for Bayesian Inference*, 2nd ed. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2260716
GAUCH JR, H. G. (2013). A simple protocol for AMMI analysis of yield trials. *Crop Sci.* **53** 1860–1869.
GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 339–373.
GOLLOB, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* **33** 73–115. MR0221658 https://doi.org/10.1007/BF02289676
GOOD, I. J. (1969). Some applications of the singular decomposition of a matrix. *Technometrics* **11** 823–831.
GOODMAN, L. A. and HABERMAN, S. J. (1990). The analysis of nonadditivity in two-way analysis of variance. *J. Amer. Statist. Assoc.* **85** 139–145. MR1137360
GU, C. (2014). Smoothing spline ANOVA models: R package gss. *J. Stat. Softw.* **58** 1–25.
GUHANIYOGI, R., QAMAR, S. and DUNSON, D. B. (2017). Bayesian tensor regression. *J. Mach. Learn. Res.* **18** 2733–2763.
HARSHMAN, R. A. and LUNDY, M. E. (1994). PARAFAC: Parallel factor analysis. *Comput. Statist. Data Anal.* **18** 39–72.
HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statist. Sci.* **15** 196–223. MR1820768 https://doi.org/10.1214/ss/1009212815
HERNÁNDEZ, B., PENNINGTON, S. R., PARNELL, A. C. et al. (2015). Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics* **9** 54–64.
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
ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147
ISIK, F., HOLLAND, J. and MALTECCA, C. (2017). Multi environmental trials. In *Genetic Data Analysis for Plant and Animal Breeding* 227–262. Springer, Berlin.
JEONG, S. and ROČKOVÁ, V. (2020). The art of BART: On flexibility of Bayesian forests. ArXiv preprint. Available at arXiv:2008.06620.
JOSSE, J., VAN EEUWIJK, F., PIEPHO, H.-P. and DENIS, J.-B. (2014). Another look at Bayesian analysis of AMMI models for genotype-environment data. *J. Agric. Biol. Environ. Stat.* **19** 240–257. MR3257913 https://doi.org/10.1007/s13253-014-0168-z
KAPELNER, A. and BLEICH, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *J. Stat. Softw.* **70** 1–40.
KINDO, B. P., WANG, H. and PEÑA, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat* **5** 119–131. MR3484083 https://doi.org/10.1002/sta4.110
LAL, R., CHANOTIYA, C., DHAWAN, S., GUPTA, P., MISHRA, A., SRIVASTAVA, S., SHUKLA, S. and MAURYA, R. (2020). Estimation of intra-specific genetic variability and half-sib family selection using AMMI (Additive Main Effects and Multiplicative Interactions) model in menthol mint (Mentha arvensis L.). *J. Med. Arom. Plant Sci.* **42** 102–113.
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
LINERO, A. R., BASAK, P., LI, Y. and SINHA, D. (2022). Bayesian Survival Tree Ensembles with Submodel Shrinkage. *Bayesian Anal.* **17** 997–1020. MR4505386 https://doi.org/10.1214/21-ba1285
LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241 https://doi.org/10.1214/09-BA404
LIU, Y., TRASKIN, M., LORCH, S. A., GEORGE, E. I. and SMALL, D. (2015). Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health Care Manage. Sci.* **18** 58–66.
LOVE, S. L., SALAIZ, T., SHAFII, B., PRICE, W. J., MOSLEY, A. R. and THORNTON, R. E. (2004). Stability of expression and concentration of ascorbic acid in North American potato germplasm. *HortScience* **39** 156–160.

MAHALINGAM, L., MAHENDRAN, S., BABU, R. C. and ATLIN, G. (2006). AMMI analysis for stability of grain yield in rice (Oryza sativa L.). *Int. J. Bot.*

MANDEL, J. (1971). A new analysis of variance model for non-additive data. *Technometrics* **13** 1–18.

MCCULLOCH, R., SPARAPANI, R., SPANBAUER, C., GRAMACY, R. and PRATOLA, M. (2020). BART: Bayesian Additive Regression Trees. R package version 2.8.

MITROVIAÃ, B., TRESKI, S., STOJAKKOVÃ, M., IVANOVIÃ, M. and BEKAVAC, G. (2012). Evaluation of experimental maize hybrids tested in multi-location trials using AMMI and GGE biplot analyses. *Turk. J. Field Crops* **17** 35–40.

NACHIT, M. M., NACHIT, G., KETATA, H., GAUCH, H. G. and ZOBEL, R. W. (1992). Use of AMMI and linear regression models to analyze genotype-environment interaction in durum wheat. *Theor. Appl. Genet.* **83** 597–601. https://doi.org/10.1007/BF00226903

ONOFRI, A. and CIRICIOFOLO, E. (2007). Using R to perform the AMMI analysis on agriculture variety trials. *R News* **7** 14–19.

PLUMMER, M. (2015). Cuts in Bayesian graphical models. *Stat. Comput.* **25** 37–43. MR3304902 https://doi.org/10.1007/s11222-014-9503-z

PRADO, E. B. and INGLIS, A. N. (2022). AMBARTI—Github repository.

PRADO, E. B., MORAL, R. A. and PARNELL, A. C. (2021). Bayesian additive regression trees with model trees. *Stat. Comput.* **31** Paper No. 20. MR4224945 https://doi.org/10.1007/s11222-021-09997-3

RAD, M. N., KADIR, M. A., RAFII, M., JAAFAR, H. Z., NAGHAVI, M. and AHMADI, F. (2013). Genotype environment interaction by AMMI and GGE biplot analysis in three consecutive generations of wheat (Triticum aestivum) under normal and drought stress conditions. *Aust. J. Crop Sci.* **7** 956.

ROBERT, C. and CASELLA, G. (2013). *Monte Carlo Statistical Methods*. Springer, Berlin.

ROČKOVÁ, V. and SAHA, E. (2019). On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics* 2839–2848. PMLR.

ROČKOVÁ, V. and VAN DER PAS, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.* **48** 2108–2131. MR4134788 https://doi.org/10.1214/19-AOS1879

RODRIGUES, P. C., MONTEIRO, A. and LOURENÇO, V. M. (2016). A robust AMMI model for the analysis of genotype-by-environment data. *Bioinformatics* **32** 58–66. https://doi.org/10.1093/bioinformatics/btv533

ROMAGOSA, I., ULLRICH, S. E., HAN, F. and HAYES, P. M. (1996). Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theor. Appl. Genet.* **93** 30–37.

SARTI, D. A. (2013). Uncertainty management through decision analysis: Applications to production optimization and uncertain demands Master's thesis Univ. São Paulo.

SARTI, D. A. (2019). The statistical paradigm: Probabilistic and multivariate analysis applied through computational simulation in the interaction between genotype x environment Ph.D. thesis Universidade de São Paulo.

SARTI, D. A., PRADO, E. B., INGLIS, A. N., DOS SANTOS, A. A., HURLEY, C. B., MORAL, R. A. and PARNELL, A. C. (2023). Supplement to "Bayesian additive regression trees for genotype by environment interaction models." https://doi.org/10.1214/22-AOAS1698SUPP

SATO, K. and TAKEDA, K. (1993). Pathogenic variation of pyrenophora teres isolates collected from Japanese and Canadian spring barley. *Rep. Inst. Resour. Biol. Sci., Okayama Univ.* **1** 147–158.

SHAFII, B. and PRICE, W. J. (1998). Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. *J. Agric. Biol. Environ. Stat.* **3** 335–345. MR1817043 https://doi.org/10.2307/1400587

SILVEIRA, L. C. I. D., KIST, V., PAULA, T. O. M. D., BARBOSA, M. H. P., PETERNELLI, L. A. and DAROS, E. (2013). AMMI analysis to evaluate the adaptability and phenotypic stability of sugarcane genotypes. *Sci. Agric.* **70** 27–32.

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

TAN, Y. V. and ROY, J. (2019). Bayesian additive regression trees and the General BART model. *Stat. Med.* **38** 5048–5069. MR4022845 https://doi.org/10.1002/sim.8347

TEAM, R. C. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

TYAGI, B., SINGH, M., SINGH, G., KUMAR, R., VERMA, A., SHARMA, I. et al. (2016). Genetic variability and AMMI bi-plot analysis in bread wheat based on multi-location trials conducted under drought conditions across agro-climatic zones of India. *Triticeae Genomics Genet.* **7**.

WRIGHT, M. N. and KÖNIG, I. R. (2019). Splitting on categorical predictors in random forests. *PeerJ* **7** e6339.

ZELDOW, B., RE, V. L. III and ROY, J. (2019). A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. *Ann. Appl. Stat.* **13** 1989–2010. MR4019164 https://doi.org/10.1214/19-AOAS1266

ZHANG, J. L. and HÄRDLE, W. K. (2010). The Bayesian additive classification tree applied to credit risk modelling. *Comput. Statist. Data Anal.* **54** 1197–1205. MR2600825 https://doi.org/10.1016/j.csda.2009.11.022

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1958–1983
https://doi.org/10.1214/22-AOAS1701

THE SCALABLE BIRTH–DEATH MCMC ALGORITHM FOR MIXED GRAPHICAL MODEL LEARNING WITH APPLICATION TO GENOMIC DATA INTEGRATION

BY NANWEI WANG[1,a], HÉLÈNE MASSAM[2,b], XIN GAO[2,c] AND LAURENT BRIOLLAIS[3,d]

[1]*Department of Mathematics and Statistics, University of New Brunswick,* [a]*nanwei.wang@unb.ca*

[2]*Department of Mathematics and Statistics, York University,* [b]*massamh@mathstat.yorku.ca,* [c]*xingao@yorku.ca*

[3]*Lunenfeld-Tanenbaum Research Institute of Sinai Health System and Dalla Lana School of Public Health (Biostatistics), University of Toronto,* [d]*laurent@lunenfeld.ca*

Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications that allow the study of biological mechanisms at an unprecedented depth and scale. A large amount of genomic data is now distributed through consortia like The Cancer Genome Atlas (TCGA), where specific types of biological information on specific type of tissue or cell are available. In cancer research the challenge is now to perform integrative analyses of high-dimensional multiomic data with the goal to better understand genomic processes that correlate with cancer outcomes, for example, elucidate gene networks that discriminate a specific cancer subgroups (cancer subtyping) or discovering gene networks that overlap across different cancer types (pan-cancer studies). In this paper we propose a novel mixed graphical model approach to analyze multiomic data of different types (continuous, discrete and count) and perform model selection by extending the birth–death MCMC (BDMCMC) algorithm initially proposed by Stephens (*Ann. Statist.* **28** (2000) 40–74) and later developed by Mohammadi and Wit (*Bayesian Anal.* **10** (2015) 109–138). Using simulations, we compare the performance of our method to the LASSO method and the standard BDMCMC method and find that our method is superior in terms of both computational efficiency and the accuracy of the model selection results. Finally, an application to the TCGA breast cancer data shows that integrating genomic information at different levels (mutation and expression data) leads to better subtyping of breast cancers.

REFERENCES

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. and NIELSEN, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16** 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

BERNSTEIN, B., STAMATOYANNOPOULOS, J., COSTELLO, J. et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28** 1045–1048.

CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70. https://doi.org/10.1038/nature11412

CANCER GENOME ATLAS RESEARCH NETWORK, WEINSTEIN, J., COLLISSON, E., MILLS, G., SHAW, K., OZENBERGER, B., ELLROTT, K., SHMULEVICH, I., SANDER, C. and STUART, J. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45** 1113–1120.

CAPPÉ, O., ROBERT, C. P. and RYDÉN, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 679–700. MR1998628 https://doi.org/10.1111/1467-9868.00409

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. MR3335095 https://doi.org/10.1093/biomet/asu051

Key words and phrases. SBDMCMC, mixed graphical models, genomic integration, TCGA.

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193 https://doi.org/10.1080/10618600.2016.1237362

COLAPRICO, A., SILVA, T., OLSEN, C., GAROFANO, L., CAVA, C., GAROLINI, D., SABEDOT, T., MALTA, T. et al. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44** e71.

CURTIS, C., SHAH, S., CHIN, S., TURASHVILI, G., RUEDA, O. et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486** 346–352.

DOBRA, A. and MOHAMMADI, R. (2018). Loglinear model selection and human mobility. *Ann. Appl. Stat.* **12** 815–845. MR3834287 https://doi.org/10.1214/18-AOAS1164

ENCODE (2011). ENCODE Project Consortium: A user's guide to the encyclopedia of DNA elements (ENCODE). *Nat. Genet.* **9** e1001046.

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

FELLINGHAUER, B., BÜHLMANN, P., RYFFEL, M., VON RHEIN, M. and REINHARDT, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Statist. Data Anal.* **64** 132–152. MR3061894 https://doi.org/10.1016/j.csda.2013.02.022

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441. https://doi.org/10.1093/biostatistics/kxm045

GAO, X. and CARROLL, R. J. (2017). Data integration with high dimensionality. *Biometrika* **104** 251–272. MR3698252 https://doi.org/10.1093/biomet/asx023

GUEDJ, M., MARISA, L., DE REYNIES, A., ORSETTI, B., SCHIAPPA, R. et al. (2012). A refined molecular taxonomy of breast cancer. *Oncogene* **486** 1196–1206.

HASLBECK, J. and WALDORP, L. (2020). mgm: Structure estimation for time-varying mixed graphical models in high-dimensional data. *J. Stat. Softw.* **93** 1–46.

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, New York. MR1419991

LEE, J. D. and HASTIE, T. J. (2015). Learning the structure of mixed graphical models. *J. Comput. Graph. Statist.* **24** 230–253. MR3328255 https://doi.org/10.1080/10618600.2014.900500

MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138. MR3420899 https://doi.org/10.1214/14-BA889

MOHAMMADI, R. and WIT, E. (2019). BDgraph: An R package for Bayesian structure learning in graphical models. *J. Stat. Softw.* **89** 1–30.

NAN, Y. and YANG, Y. (2014). Variable selection diagnostics measures for high-dimensional regression. *J. Comput. Graph. Statist.* **23** 636–656. MR3224649 https://doi.org/10.1080/10618600.2013.829780

PARKER, J., MULLINS, M., CHEANG, M., LEUNG, S., VODUC, D., VICKERY, T. et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27** 1160–1167.

PEROU, C., JEFFREY, S., VAN DE RIJN, M., REES, C., EISEN, M., ROSS, D. et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96** 9212–9217.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

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

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SORLIE, T., PEROU, C., TIBSHIRANI, R., AAS, T., GEISLER, S., JOHNSEN, H. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98** 10869–10874.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903 https://doi.org/10.1214/aos/1016120364

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

TURNER, N. and REIS-FILHO, J. (2013). Tackling the diversity of triple negative breast cancer. *Clin. Cancer Res.* **19** 6380–6388.

WANG, N., MASSAM, H., GAO, X. and BRIOLLAIS, L. (2023). Supplement to "The scalable birth–death MCMC algorithm for mixed graphical model learning with application to genomic data integration." https://doi.org/10.1214/22-AOAS1701SUPPA, https://doi.org/10.1214/22-AOAS1701SUPPB

WASSERMAN, L. (2000). Bayesian model selection and model averaging. *J. Math. Psych.* **44** 92–107. MR1770003 https://doi.org/10.1006/jmps.1999.1278

WEIGELT, B., PUSZTAI, L., ASHWORTH, A. and REIS-FILHO, J. S. (2011). Challenges translating breast cancer gene signatures into the clinic. *Nat. Rev. Clin. Oncol.* **9** 58–64. https://doi.org/10.1038/nrclinonc.2011.125

YANG, E., BAKER, Y., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2014). Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics* 1042–1050.

YE, C., YANG, Y. and YANG, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *J. Amer. Statist. Assoc.* **113** 1797–1812. MR3902247 https://doi.org/10.1080/01621459.2017.1377080

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

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 1984–2001
https://doi.org/10.1214/22-AOAS1703

A BAYESIAN HIERARCHICAL MODEL FRAMEWORK TO QUANTIFY UNCERTAINTY OF TROPICAL CYCLONE PRECIPITATION FORECASTS

BY STEPHEN WALSH[1,a], MARCO A. R. FERREIRA[1,b], DAVID HIGDON[1,c] AND STEPHANIE ZICK[2,d]

[1]*Department of Statistics, Virginia Tech,* [a]*walsh124@vt.edu,* [b]*marf@vt.edu,* [c]*dhigdon@vt.edu*
[2]*Department of Geography, Virginia Tech,* [d]*sezick@vt.edu*

Tropical cyclones present a serious threat to many coastal communities around the world. Many numerical weather prediction models provide deterministic forecasts with limited measures of their forecast uncertainty. Standard postprocessing techniques may struggle with extreme events or use a 30-day training window that will not adequately characterize the uncertainty of a tropical cyclone forecast. We propose a novel approach that leverages information from past storm events, using a hierarchical model to quantify uncertainty in the spatial correlation parameters of the forecast errors (modeled as Gaussian processes) for a numerical weather prediction model. This approach addresses a massive data problem by implementing a drastic dimension reduction through the assumption that the MLE and Hessian matrix represent all useful information from each tropical cyclone. From this, simulated forecast errors provide uncertainty quantification for future tropical cyclone forecasts. We apply this method to the North American Mesoscale model forecasts and use observations based on the Stage IV data product for 47 tropical cyclones between 2004 and 2017. For an incoming storm, our hierarchical framework combines the forecast from the North American Mesoscale model with the information from previous storms to create 95% and 99% prediction maps of rain. For six test storms from 2018 and 2019, these maps provide appropriate probabilistic coverage of observations. We show evidence from the log scoring rule that the proposed hierarchical framework performs best among competing methods.

REFERENCES

ACCADIA, C., MARIANI, S., CASAIOLI, M., LAVAGNINI, A. and SPERANZA, A. (2003). Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather Forecast*. **18** 918–932.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

BANNISTER, R. (2017). A review of operational methods of variational and ensemble-variational data assimilation. *Q. J. R. Meteorol. Soc*. **143** 607–633.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2007). Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Weather Rev.* **135** 1386–1402.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* **2** 1170–1193. MR2655654 https://doi.org/10.1214/08-AOAS203

BHATIA, K. T., VECCHI, G. A., KNUTSON, T. R., MURAKAMI, H., KOSSIN, J., DIXON, K. W. and WHITLOCK, C. E. (2019). Recent increases in tropical cyclone intensification rates. *Nat. Commun*. **10** 1–9.

BISHOP, C. H. and SHANLEY, K. T. (2008). Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Weather Rev.* **136** 4641–4652.

BORCHERS, H. W. (2019). pracma: Practical Numerical Math Functions. R package version 2.2.5.

Key words and phrases. Bayesian statistics, hurricane forecasts, massive datasets, meteorology, spatial statistics, uncertainty quantification.

CLARK, A. J., GALLUS, W. A. JR. and WEISMAN, M. L. (2010). Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Weather Forecast*. **25** 1495–1509.

CUTTER, S. L., EMRICH, C. T., MITCHELL, J. T., PIEGORSCH, W. W., SMITH, M. M. and WEBER, L. (2014). *Hurricane Katrina and the Forgotten Coast of Mississippi*. Cambridge Univ. Press, Cambridge.

DIGGLE, P. J. and RIBEIRO, P. J. JR. (2007). *Model-Based Geostatistics*. *Springer Series in Statistics*. Springer, New York. MR2293378

FELDMANN, K., SCHEUERER, M. and THORARINSDOTTIR, T. L. (2015). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Weather Rev.* **143** 955–971.

GEL, Y., RAFTERY, A. E. and GNEITING, T. (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method. *J. Amer. Statist. Assoc.* **99** 575–583. MR2086380 https://doi.org/10.1198/016214504000000872

GNEITING, T. and KATZFUSS, M. (2014). Probabilistic forecasting. *Annu. Rev. Stat. Appl.* **1** 125–151.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

GNEITING, T., RAFTERY, A. E., WESTVELD, A. H. III and GOLDMAN, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133** 1098–1118.

GOOD, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. Ser. B* **14** 107–114. MR0077033

HABIB, E., HENSCHKE, A. and ADLER, R. F. (2009). Evaluation of TMPA satellite-based research and real-time rainfall estimates during six tropical-related heavy rainfall events over Louisiana, USA. *Atmos. Res.* **94** 373–388.

HABIB, E., LARSON, B. F. and GRASCHEL, J. (2009). Validation of NEXRAD multisensor precipitation estimates using an experimental dense rain gauge network in South Louisiana. *J. Hydrol.* **373** 463–478.

HAMILL, T. M., BRENNAN, M. J., BROWN, B., DEMARIA, M., RAPPAPORT, E. N. and TOTH, Z. (2012). NOAA's future ensemble-based hurricane forecast products. *Bull. Am. Meteorol. Soc.* **93** 209–220.

JAGGER, T. H. and ELSNER, J. B. (2006). Climatology models for extreme hurricane winds near the United States. *J. Climate* **19** 3220–3236.

JANJIC, Z. (2003). A nonhydrostatic model based on a new approach. *Meteorol. Atmos. Phys.* **82** 271–285.

JIANG, H., HALVERSON, J. B. and SIMPSON, J. (2008). On the differences in storm rainfall from Hurricanes Isidore and Lili. Part I: Satellite observations and rain potential. *Weather Forecast.* **23** 29–43.

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 https://doi.org/10.1111/1467-9868.00294

KLEIBER, W., RAFTERY, A. E. and GNEITING, T. (2011). Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Statist. Assoc.* **106** 1291–1303. MR2896836 https://doi.org/10.1198/jasa.2011.ap10433

KLOTZBACH, P. J., BELL, M. M. and JONES, J. (2020). Summary of 2020 Atlantic Tropical Cyclone Activity and Verification of Authors' Seasonal and Two-Week Forecasts Technical Report Colorado State Univ.

KO, M.-C., MARKS, F. D., ALAKA, G. J. and GOPALAKRISHNAN, S. G. (2020). Evaluation of hurricane Harvey (2017) rainfall in deterministic and probabilistic HWRF forecasts. *Atmosphere* **11** 666.

KRZYSZTOFOWICZ, R. and EVANS, W. B. (2008). Probabilistic forecasts from the national digital forecast database. *Weather Forecast.* **23** 270–289.

LANDSEA, C. W. and FRANKLIN, J. L. (2013). Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Weather Rev.* **141** 3576–3592.

LEWIS, S. M. and RAFTERY, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *J. Amer. Statist. Assoc.* **92** 648–655. MR1467855 https://doi.org/10.2307/2965712

LI, W., DUAN, Q., MIAO, C., YE, A., GONG, W. and DI, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews*: *Water* **4** e1246.

LIN, Y. and MITCHELL, K. E. (2005). 1.2 the NCEP stage II/IV hourly precipitation analyses: Development and applications. In 19*th Conf. Hydrology*, *American Meteorological Society*, *San Diego*, *CA*, *USA* 1–4. Citeseer.

LORENZ, E. N. (1996). Predictability: A problem partly solved. In *Proc. Seminar on Predictability* **1** 1–18. ECMWF.

MARCHOK, T., ROGERS, R. and TULEYA, R. (2007). Validation schemes for tropical cyclone quantitative precipitation forecasts: Evaluation of operational models for U.S. landfalling cases. *Weather Forecast.* **22** 726–746.

NELSON, B. R., PRAT, O. P., SEO, D.-J. and HABIB, E. (2016). Assessment and implications of NCEP stage IV quantitative precipitation estimates for product intercomparisons. *Weather Forecast.* **31** 371–394.

RAFTERY, A. E., GNEITING, T., BALABDAOUI, F. and POLAKOWSKI, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133** 1155–1174.

ROGERS, E., DIMEGO, G., BLACK, T., EK, M., FERRIER, B., GAYNO, G., JANJIC, Z., LIN, Y., PYLE, M. et al. (2009). The NCEP North American mesoscale modeling system: Recent changes and future plans. In 23*rd Conf. on Weather Analysis and Forecasting*/19*th Conf. on Numerical Weather Prediction*, *Omaha*, *NE*, *Amer. Meteor. Soc*., 2*A*.4.

SCHAAKE, J., DEMARGNE, J., HARTMAN, R., MULLUSKY, M., WELLES, E., WU, L., HERR, H., FAN, X. and SEO, D. (2007). Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci. Discuss.* **4** 655–717.

SCHUURMAN, N. K., GRASMAN, R. P. P. P. and HAMAKER, E. L. (2016). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivar. Behav. Res.* **51** 185–206. https://doi.org/10.1080/00273171.2015.1065398

SLOUGHTER, J. M. L., RAFTERY, A. E., GNEITING, T. and FRALEY, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Weather Rev.* **135** 3209–3220.

VILLARINI, G., SMITH, J. A., BAECK, M. L., MARCHOK, T. and VECCHI, G. A. (2011). Characterization of rainfall distribution and flooding associated with US landfalling tropical cyclones: Analyses of Hurricanes Frances, Ivan, and Jeanne (2004). *J. Geophys. Res., Atmos.* **116** 1–19.

VILLARINI, G., ZHANG, W., MILLER, P., JOHNSON, D. R., GRIMLEY, L. E. and ROBERTS, H. J. (2022). Probabilistic rainfall generator for tropical cyclones affecting Louisiana. *Int. J. Climatol.* **42** 1789–1802.

WALSH, S., FERREIRA, M. A., HIGDON, D. and ZICK, S. (2023a). Supplement to "A Bayesian hierarchical model framework to quantify uncertainty of tropical cyclone precipitation forecasts." https://doi.org/10.1214/22-AOAS1703SUPPA

WALSH, S., FERREIRA, M. A., HIGDON, D. and ZICK, S. (2023b). Supplement to "A Bayesian hierarchical model framework to quantify uncertainty of tropical cyclone precipitation forecasts." https://doi.org/10.1214/22-AOAS1703SUPPB

WILLIAMS, R., FERRO, C. and KWASNIOK, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Q. J. R. Meteorol. Soc.* **140** 1112–1120.

YAN, H. and GALLUS, W. A. JR. (2016). An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain. *Weather Forecast.* **31** 1363–1379.

ZAGRODNIK, J. P. and JIANG, H. (2013). Investigation of PR and TMI version 6 and version 7 rainfall algorithms in landfalling tropical cyclones relative to the NEXRAD stage-IV multisensor precipitation estimate dataset. *J. Appl. Meteorol. Climatol.* **52** 2809–2827.

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

ZICK, S. E. and MATYAS, C. J. (2016). A shape metric methodology for studying the evolving geometries of synoptic-scale precipitation patterns in tropical cyclones. *Ann. Assoc. Amer. Geogr.* **106** 1217–1235.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2002–2018
https://doi.org/10.1214/22-AOAS1704

JOINT POINT AND VARIANCE ESTIMATION UNDER A HIERARCHICAL BAYESIAN MODEL FOR SURVEY COUNT DATA

BY TERRANCE D. SAVITSKY[1,a], JULIE GERSHUNSKAYA[2,b] AND MARK CRANKSHAW[2,c]

[1]*Office of Survey Methods Research, U.S. Bureau of Labor Statistics,* [a]*Savitsky.Terrance@bls.gov*
[2]*OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics,* [b]*Gershunskaya.Julie@bls.gov,* [c]*Crankshaw.Mark@bls.gov*

We propose a novel Bayesian framework for the joint modeling of survey point and variance estimates for count data. The approach incorporates an induced prior distribution on the modeled true variance that sets it equal to the generating variance of the point estimate, a key property more readily achieved for continuous data response type models. Our count data model formulation allows the input of domains at multiple resolutions (e.g., states, regions, nation) and simultaneously benchmarks modeled estimates at higher resolutions (e.g., states) to those at lower resolutions (e.g., regions) in a fashion that borrows more strength to sharpen our domain estimates at higher resolutions. We conduct a simulation study that generates a population of units within domains to produce ground truth statistics to compare to direct and modeled estimates performed on samples taken from the population where we show improved reductions in error across domains. The model is applied to the job openings variable and other data items published in the Job Openings and Labor Turnover Survey administered by the U.S. Bureau of Labor Statistics.

REFERENCES

BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American community survey. *J. Amer. Statist. Assoc.* **111** 472–487. MR3538680 https://doi.org/10.1080/01621459.2015.1117471

GELMAN, A., LEE, D. and GUO, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *J. Educ. Behav. Stat.* **40** 530–543. https://doi.org/10.3102/1076998615606113

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

MAITI, T., REN, H. and SINHA, S. (2014). Prediction error of small area predictors shrinking both means and variances. *Scand. J. Stat.* **41** 775–790. MR3249428 https://doi.org/10.1111/sjos.12061

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

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. MR3380626 https://doi.org/10.1002/9781118735855

SAVITSKY, T. D. (2016). Bayesian nonparametric multiresolution estimation for the American Community Survey. *Ann. Appl. Stat.* **10** 2157–2181. MR3592052 https://doi.org/10.1214/16-AOAS968

SAVITSKY, T. D., GERSHUNSKAYA, J. and CRANKSHAW, M. (2023). Supplement to "Joint point and variance estimation under a hierarchical Bayesian model for survey count data." https://doi.org/10.1214/22-AOAS1704SUPP

SUGASAWA, S. and KUBOKAWA, T. (2020). Small area estimation with mixed models: A review. *Jpn. J. Stat. Data Sci.* **3** 693–720. MR4181996 https://doi.org/10.1007/s42081-020-00076-x

SUGASAWA, S., TAMAE, H. and KUBOKAWA, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scand. J. Stat.* **44** 150–167. MR3619699 https://doi.org/10.1111/sjos.12246

TZAVIDIS, N., RANALLI, M. G., SALVATI, N., DREASSI, E. and CHAMBERS, R. (2015). Robust small area prediction for counts. *Stat. Methods Med. Res.* **24** 373–395. MR3350290 https://doi.org/10.1177/0962280214520731

Key words and phrases. Bayesian hierarchical models, small area estimation, count data, Stan.

WEST, M. (2013). Bayesian dynamic modelling. In *Bayesian Theory and Applications* (P. Damien, P. Dellaportas, D. A. Polson and N. G. Stephens, eds.) 145–166. Oxford Univ. Press, Oxford. MR3221162

ZHOU, M., LI, L., DUNSON, D. and CARIN, L. (2012). Lognormal and gamma mixed. In *Negative Binomial Regression*. *Proc*. *Int*. *Conf*. *Mach*. *Learn*. **2012** 1343–1350. PMID: 25279391; PMCID: PMC4180062..

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2019–2038
https://doi.org/10.1214/22-AOAS1705

DATA-ADAPTIVE DISCRIMINATIVE FEATURE LOCALIZATION WITH STATISTICALLY GUARANTEED INTERPRETATION

BY BEN DAI[1,a], XIAOTONG SHEN[2,b], LIN YEE CHEN[3,d], CHUNLIN LI[2,c] AND WEI PAN[4,e]

[1]*Department of Statistics, The Chinese University of Hong Kong,* [a]*bendai@cuhk.edu.hk*
[2]*School of Statistics, University of Minnesota,* [b]*xshen@umn.edu,* [c]*li000007@umn.edu*
[3]*Lillehei Heart Institute and Cardiovascular Division, University of Minnesota,* [d]*chenx484@umn.edu*
[4]*Division of Biostatistics, University of Minnesota,* [e]*panxx014@umn.edu*

In explainable artificial intelligence, discriminative feature localization is critical to reveal a black-box model's decision-making process from raw data to prediction. In this article we use two real datasets, the MNIST handwritten digits and MIT-BIH electrocardiogram (ECG) signals, to motivate key characteristics of discriminative features, namely, *adaptiveness*, *predictive importance* and *effectiveness*. Then we develop a localization framework, based on adversarial attacks, to effectively localize discriminative features. In contrast to existing heuristic methods, we also provide a statistically guaranteed interpretability of the localized features by measuring a generalized partial R^2. We apply the proposed method to the MNIST dataset and the MIT-BIH dataset with a convolutional autoencoder. In the first, the compact image regions localized by the proposed method are visually appealing. Similarly, in the second, the identified ECG features are biologically plausible and consistent with cardiac electrophysiological principles while locating subtle anomalies in a QRS complex that may not be discernible by the naked eye. Overall, the proposed method compares favorably with state-of-the-art competitors. Accompanying this paper is a Python library **dnn-locate** that implements the proposed approach.

REFERENCES

ACHARYA, U. R., OH, S. L., HAGIWARA, Y., TAN, J. H., ADAM, M., GERTYCH, A. and SAN TAN, R. (2017). A deep convolutional neural network model to classify heartbeats. *Comput. Biol. Med.* **89** 389–396.

ATTIA, Z. I., NOSEWORTHY, P. A., LOPEZ-JIMENEZ, F., ASIRVATHAM, S. J., DESHMUKH, A. J., GERSH, B. J., CARTER, R. E., YAO, X., RABINSTEIN, A. A. et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet* **394** 861–867.

BACH, S., BINDER, A., MONTAVON, G., KLAUSCHEN, F., MÜLLER, K.-R. and SAMEK, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10** e0130140. https://doi.org/10.1371/journal.pone.0130140

BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554 https://doi.org/10.1214/009053605000000282

BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032 https://doi.org/10.1198/016214505000000907

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026 https://doi.org/10.1162/153244303321897690

BENGIO, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural Networks*: *Tricks of the Trade* 437–478. Springer, Berlin.

BERGSTRA, J. and BENGIO, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13** 281–305. MR2913701

BHARTI, R., KHAMPARIA, A., SHABAZ, M., DHIMAN, G., PANDE, S. and SINGH, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Comput. Intell. Neurosci.* **2021** 8387680. https://doi.org/10.1155/2021/8387680

Key words and phrases. Explainable artificial intelligence, discriminative features, localization, generalized partial R^2, interpretability, regularization, deep learning.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.
DAI, B., SHEN, X., CHEN, L. Y, LI, C. and PAN, W. (2023). Supplement to "Data-adaptive discriminative feature localization with statistically guaranteed interpretation." https://doi.org/10.1214/22-AOAS1705SUPPA, https://doi.org/10.1214/22-AOAS1705SUPPB
DAI, B., SHEN, X. and PAN, W. (2022). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.*
DAVIS, R. A., LII, K.-S. and POLITIS, D. N. (2011). Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt* 95–100. Springer, Berlin.
ELGENDI, M. (2013). Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases. *PLoS ONE* **8** e73557. https://doi.org/10.1371/journal.pone.0073557
EVANS, R., O'NEILL, M., PRITZEL, A., ANTROPOVA, N., SENIOR, A. W., GREEN, T., ŽÍDEK, A., BATES, R., BLACKWELL, S. et al. (2021). Protein complex prediction with AlphaFold-Multimer. *Biorxiv*.
FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451
GE, R., HUANG, F., JIN, C. and YUAN, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory* 797–842.
GHORBANI, A., ABID, A. and ZOU, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 3681–3688.
HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.
HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Comput.* **9** 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
JAMBUKIA, S. H., DABHI, V. K. and PRAJAPATI, H. B. (2015). Classification of ECG signals using machine learning techniques: A survey. In 2015 *International Conference on Advances in Computer Engineering and Applications* 714–721. IEEE, Ghaziabad.
JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596** 583–589.
KACHUEE, M., FAZELI, S. and SARRAFZADEH, M. (2018). Ecg heartbeat classification: A deep transferable representation. In 2018 *IEEE International Conference on Healthcare Informatics* (*ICHI*) 443–444. IEEE, New York.
KINDERMANS, P.-J., SCHÜTT, K. T., ALBER, M., MÜLLER, K.-R., ERHAN, D., KIM, B. and DÄHNE, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. ArXiv preprint. Available at arXiv:1705.05598.
KO, W.-Y., SIONTIS, K. C., ATTIA, Z. I., CARTER, R. E., KAPA, S., OMMEN, S. R., DEMUTH, S. J., ACKERMAN, M. J., GERSH, B. J. et al. (2020). Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J. Am. Coll. Cardiol.* **75** 722–733.
KUSUMOTO, F. (2020). *ECG Interpretation*: *From Pathophysiology to Clinical Application*. Springer Nature, Berlin.
LECUN, Y. and CORTES, C. (2010). MNIST handwritten digit database.
LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W. and JACKEL, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1** 541–551.
LEE, J. D., SIMCHOWITZ, M., JORDAN, M. I. and RECHT, B. (2016). Gradient descent only converges to minimizers. In *Conference on Learning Theory* 1246–1257.
LIN, Y. (2004). A note on margin-based loss functions in classification. *Statist. Probab. Lett.* **68** 73–82. MR2064687 https://doi.org/10.1016/j.spl.2004.03.002
LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774.
MARTIS, R. J., ACHARYA, U. R., LIM, C. M., MANDANA, K., RAY, A. K. and CHAKRABORTY, C. (2013). Application of higher order cumulant features for cardiac health diagnosis using ECG signals. *Int. J. Neural Syst.* **23** 1350014.
MASCI, J., MEIER, U., CIREŞAN, D. and SCHMIDHUBER, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks* 52–59. Springer, Berlin.
MCFADDEN, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.
MIRVIS, D. M. and GOLDBERGER, A. L. (2001). Electrocardiography. In *Heart Disease* 82–128. W. B. Saunders, Philadelphia.
MONTAVON, G., LAPUSCHKIN, S., BINDER, A., SAMEK, W. and MÜLLER, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **65** 211–222.

MOODY, G. B. and MARK, R. G. (1990). The MIT-BIH arrhythmia database on CD-ROM and software for use with it. In [1990] *Proceedings Computers in Cardiology* 185–188. IEEE, Chicago.

NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78** 691–692. MR1130937 https://doi.org/10.1093/biomet/78.3.691

RAGINSKY, M., RAKHLIN, A. and TELGARSKY, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. ArXiv preprint. Available at arXiv:1702.03849.

RAJPURKAR, P., HANNUN, A. Y., HAGHPANAHI, M., BOURN, C. and NG, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. ArXiv preprint. Available at arXiv:1707.01836.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.* **15** 335–366. MR3190843

RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144.

RUDY, Y. (2004). Ionic mechanisms of cardiac electrical activity: A theoretical approach. In *Cardiac Electrophysiology*: *From Cell to Bedside* 255–266. Elsevier, Philadelphia.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1985). Learning internal representations by error propagation Technical Report California Univ. San Diego La Jolla Inst. for Cognitive Science.

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. and BATRA, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626.

STERGIOU, G. S., ALPERT, B., MIEKE, S., ASMAR, R., ATKINS, N., ECKERT, S., FRICK, G., FRIEDMAN, B., GRASSL, T. et al. (2018). A universal standard for the validation of blood pressure measuring devices: Association for the advancement of medical instrumentation/European society of hypertension/international organization for standardization (AAMI/ESH/ISO) collaboration statement. *J. Hypertens.* **71** 368–374.

THYGESEN, K., ALPERT, J. S., WHITE, H. D. and JOINT ESC/ACCF/AHA/WHF TASK FORCE FOR THE REDEFINITION OF MYOCARDIAL INFARCTION (2007). Universal definition of myocardial infarction. *J. Am. Coll. Cardiol.* **50** 2173–2195.

TJOA, E. and GUAN, C. (2019). A survey on explainable artificial intelligence (XAI): Towards medical XAI. ArXiv preprint. Available at arXiv:1907.07374.

VAMATHEVAN, J., CLARK, D., CZODROWSKI, P., DUNHAM, I., FERRAN, E., LEE, G., LI, B., MADABHUSHI, A., SHAH, P. et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18** 463–477.

WANG, H., WANG, N. and YEUNG, D.-Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1235–1244.

WASIMUDDIN, M., ELLEITHY, K., ABUZNEID, A.-S., FAEZIPOUR, M. and ABUZAGHLEH, O. (2020). Stages-based ECG signal analysis from traditional signal processing to machine learning approaches: A survey. *IEEE Access* **8** 177782–177803.

WU, Y. and LIU, Y. (2007). Robust truncated hinge loss support vector machines. *J. Amer. Statist. Assoc.* **102** 974–983. MR2411659 https://doi.org/10.1198/016214507000000617

ZEILER, M. D. and FERGUS, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* 818–833. Springer, Berlin.

ZHOU, B., KHOSLA, A., LAPEDRIZA, A., OLIVA, A. and TORRALBA, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2921–2929.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2039–2058
https://doi.org/10.1214/22-AOAS1706

DYNAMIC PREDICTION OF RESIDUAL LIFE WITH LONGITUDINAL COVARIATES USING LONG SHORT-TERM MEMORY NETWORKS

BY GRACE RHODES[a], MARIE DAVIDIAN[b] AND WENBIN LU[c]

Department of Statistics, North Carolina State University, [a]*gmrhodes@ncsu.edu,* [b]*davidian@ncsu.edu,* [c]*wlu4@ncsu.edu*

Sepsis, a complex medical condition that involves severe infections with life-threatening organ dysfunction, is a leading cause of death worldwide. Treatment of sepsis is highly challenging. When making treatment decisions, clinicians and patients desire accurate predictions of mean residual life (MRL) that leverage all available patient information, including longitudinal biomarker data. Biomarkers are biological, clinical, and other variables reflecting disease progression that are often measured repeatedly on patients in the clinical setting. Dynamic prediction methods leverage accruing biomarker measurements to improve performance, providing updated predictions as new measurements become available. We introduce two methods for dynamic prediction of MRL using longitudinal biomarkers. In both methods, we begin by using long short-term memory networks (LSTMs) to construct encoded representations of the biomarker trajectories, referred to as "context vectors." In our first method, the LSTM-GLM, we dynamically predict MRL via a transformed MRL model that includes the context vectors as covariates. In our second method, the LSTM-NN, we dynamically predict MRL from the context vectors using a feed-forward neural network. We demonstrate the improved performance of both proposed methods relative to competing methods in simulation studies. We apply the proposed methods to dynamically predict the restricted mean residual life (RMRL) of septic patients in the intensive care unit using electronic medical record data. We demonstrate that the LSTM-GLM and the LSTM-NN are useful tools for producing individualized, real-time predictions of RMRL that can help inform the treatment decisions of septic patients.

REFERENCES

AGGARWAL, C. C. (2018). *Neural Networks and Deep Learning*. Springer, Cham. MR3966422 https://doi.org/10.1007/978-3-319-94463-0

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48. https://doi.org/10.18637/jss.v067.i01

CARRARA, M., BASELLI, G. and FERRARIO, M. (2015). Mortality prediction model of septic shock patients based on routinely recorded data. *Comput. Math. Methods Med.* **2015**. https://doi.org/10.1155/2015/761435

CHEN, Y. Q. (2007). Additive expectancy regression. *J. Amer. Statist. Assoc.* **102** 153–166. MR2345536 https://doi.org/10.1198/016214506000000870

CHOLLET, F. et al. (2015). Keras. https://keras.io.

EVANS, L., RHODES, A., ALHAZZANI, W., ANTONELLI, M., COOPERSMITH, C. M., FRENCH, C., MACHADO, F. R., MCINTYRE, L., OSTERMANN, M. et al. (2021). Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Crit. Care Med.* **49** e1063–e1143. https://doi.org/10.1097/CCM.0000000000005337

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

GAJARDO, A., BHATTACHARJEE, S., CARROLL, C., CHEN, Y., DAI, X., FAN, J., HADJIPANTELIS, P. Z., HAN, K., JI, H. et al. (2021). fdapace: Functional data analysis and empirical dynamics. R package version 0.5.8.

Key words and phrases. Biomarker, dynamic prediction, electronic medical record, long short-term memory network, longitudinal data, MIMIC-III, neural network, residual life, sepsis, transformed mean residual life model.

HARRELL, F. E., LEE, K. L. and MARK, D. B. (1996). Tutorial in biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387. https://doi.org/10.1002/0470023678.ch2b(i)

HICKEY, G. L., PHILIPSON, P., JORGENSEN, A. and KOLAMUNNAGE-DONA, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Med. Res. Methodol.* **16** 1–15. https://doi.org/10.1186/s12874-016-0212-5

HOU, N., LI, M., HE, L., XIE, B., WANG, L., ZHANG, R., YU, Y., SUN, X., PAN, Z. et al. (2020). Predicting 30-days mortality for MIMIC-III patients with Sepsis-3: A machine learning approach using XGboost. *J. Transl. Med.* **18** 1–14. https://doi.org/10.1186/s12967-020-02620-5

JOHNSON, A. E., POLLARD, T. J., SHEN, L., LEHMAN, L. H., FENG, M., GHASSEMI, M., MOODY, B., SZOLOVITS, P., CELI, L. A. et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* **3** 160035.

KINGMA, D. P. and BA, J. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG].

KOMOROWSKI, M. (2019). AI Clinician. GitHub repository. Available at https://github.com/matthieukomorowski/AI_Clinician.

LÁSZLÓ, I., TRÁSY, D., MOLNÁR, Z. and FAZAKAS, J. (2015). Sepsis: From pathophysiology to individualized patient care. *J. Immunol. Res.* **2015**. https://doi.org/10.1155/2015/510436

LIN, X., LU, T., YAN, F., LI, R. and HUANG, X. (2018). Mean residual life regression with functional principal component analysis on longitudinal data for dynamic prediction. *Biometrics* **74** 1482–1491. MR3908164 https://doi.org/10.1111/biom.12876

MAGULURI, G. and ZHANG, C.-H. (1994). Estimation in the mean residual life regression model. *J. Roy. Statist. Soc. Ser. B* **56** 477–489. MR1278221

O'MALLEY, T., BURSZTEIN, E., LONG, J., CHOLLET, F., JIN, H., INVERNIZZI, L. et al. (2019). KerasTuner. https://github.com/keras-team/keras-tuner.

RHODES, G., DAVIDIAN, M. and LU, W. (2023). Supplement to "Dynamic prediction of residual life with longitudinal covariates using long short-term memory networks." https://doi.org/10.1214/22-AOAS1706SUPPA, https://doi.org/10.1214/22-AOAS1706SUPPB

RIZOPOULOS, D., MOLENBERGHS, G. and LESAFFRE, E. M. E. H. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biom. J.* **59** 1261–1276. MR3731215 https://doi.org/10.1002/bimj.201600238

SINGER, M., DEUTSCHMAN, C. S., SEYMOUR, C. W., SHANKAR-HARI, M., ANNANE, D., BAUER, M., BELLOMO, R., BERNARD, G. R., CHICHE, J.-D. et al. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *J. Amer. Med. Assoc.* **315** 801–810. https://doi.org/10.1001/jama.2016.0287

STRIMBU, K. and TAVEL, J. A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS* **5** 463–6. https://doi.org/10.1097/COH.0b013e32833ed177

SUN, L., SONG, X. and ZHANG, Z. (2012). Mean residual life models with time-dependent coefficients under right censoring. *Biometrika* **99** 185–197. MR2899672 https://doi.org/10.1093/biomet/asr065

SUN, L. and ZHANG, Z. (2009). A class of transformed mean residual life models with censored survival data. *J. Amer. Statist. Assoc.* **104** 803–815. MR2541596 https://doi.org/10.1198/jasa.2009.0130

THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health*. Springer, New York. MR1774977 https://doi.org/10.1007/978-1-4757-3294-8

TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

VAN HOUWELINGEN, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scand. J. Stat.* **34** 70–85. MR2325243 https://doi.org/10.1111/j.1467-9469.2006.00529.x

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

WERBOS, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proc. IEEE* **78** 1550–1560. https://doi.org/10.1109/5.58337

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZHENG, Y. and HEAGERTY, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics* **61** 379–391. MR2140909 https://doi.org/10.1111/j.1541-0420.2005.00323.x

ZHU, Y., LI, L. and HUANG, X. (2019). Landmark linear transformation model for dynamic prediction with application to a longitudinal cohort study of chronic disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 771–791. MR3937473

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2059–2077
https://doi.org/10.1214/22-AOAS1707

POSTELECTION ANALYSIS OF PRESIDENTIAL ELECTION/POLL DATA

BY JIMING JIANG[1,a], YUANYUAN LI[1,b] AND PETER X. K. SONG[2,c]

[1]*Department of Statistics, University of California, Davis,* [a]*jimjiang@ucdavis.edu,* [b]*yynli@ucdavis.edu*
[2]*Department of Biostatistics, University of Michigan,* [c]*pxsong@umich.edu*

This paper concerns analyses of the 2016 and 2020 U.S. presidential election data, including the data of preelection polls and the actual elections. Our analyses unveil statistical evidence of discrepancy between the polls and real elections that is consistent across these two elections. Specifically, the polls had consistently overestimated advantages of the Democratic candidates or, equivalently, underestimated the true population support of the Republican candidate, Donald Trump, in both elections. The analyses are stratified by state, reflecting the U.S. electoral college system by the means of small area estimation. We have found recurrent patterns suggesting that the polls have been underestimating the Republican candidate, especially in swing states of critical importance. Our findings also suggest an improvement of the 2020 polling methods to mitigate the size of underestimation. We show that a small-area model built upon the actual election data from one election can provide a better prediction than the poll-based projection to another election involving the same Republican candidate. Ranking of pollsters, based on prediction bias, using mixed model prediction is also considered.

REFERENCES

CHESNEY, T. and PENNY, K. (2013). The impact of repeated lying on survey results. *SAGE Open* **3** 2158244012472345. https://doi.org/10.1177/2158244012472345

DI BRISCO, A. M. and MIGLIORATI, S. (2021). A spatial mixed-effects regression model for electoral data. *Stat. Methods Appl.* **30** 543–571. MR4266411 https://doi.org/10.1007/s10260-020-00534-6

FELSENTHAL, D. S., MAOZ, Z. and RAPOPORT, A. (1993). An empirical evaluation of six voting procedures: Do they really make any difference? *Br. J. Polit. Sci.* **23** 1–27.

FEREJOHN, J. and FIORINA, M. (1974). The paradox of not voting: A decision theoretic analysis. *Amer. Polit. Sci. Rev.* **68** 525–536.

GELMAN, A. and KING, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *Br. J. Polit. Sci.* **23** 409–451.

GELMAN, A. and KING, G. (1994). A unified method of evaluating electoral systems and redistricting plans. *Amer. J. Polit. Sci.* **38** 514–554.

GELMAN, A. and LITTLE, T. (1997). Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* **23** 127–135.

JIANG, J. and LAHIRI, P. (2006). Mixed model prediction and small area estimation. *TEST* **15** 1–96. MR2252522 https://doi.org/10.1007/BF02595419

JIANG, J., LI, Y. and SONG, P. X. (2023). Supplement to "Postelection analysis of presidential election/poll data." https://doi.org/10.1214/22-AOAS1707SUPPA, https://doi.org/10.1214/22-AOAS1707SUPPB

JIANG, J. and NGUYEN, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2308058

JIANG, J. and TORABI, M. (2020). Sumca: Simple, unified, Monte-Carlo-assisted approach to second-order unbiased mean-squared prediction error estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 467–485. MR4084172

KATZ, J. N., GELMAN, A. and KING, G. (2002). Empirically evaluating the electoral college. In *Rethinking the Vote*: *The Politics and Prospects of American Election Reform* (A. N. Crigler, M. R. Just and E. J. McCaffery, eds.) Oxford Univ. Press, London.

LINZER, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *J. Amer. Statist. Assoc.* **108** 124–134. MR3174607 https://doi.org/10.1080/01621459.2012.737735

Key words and phrases. Empirical BLUP, measure of uncertainty, mixed-effects model, opinion polls, projection, small area estimation.

MERRILL, S. III (1978). Citizen voting power under the electoral college: A stochastic model based on state voting patterns. *SIAM J. Appl. Math.* **34** 376–390. MR0475996 https://doi.org/10.1137/0134031
PARK, D., GELMAN, A. and BAFUMI, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Polit. Anal.* **12** 375–385.
PRATT, L. Y. and THRUN, S. (1997). *Machine Learning—Special Issue on Inductive Transfer*. Springer, Berlin.
RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. MR3380626 https://doi.org/10.1002/9781118735855
RUSK, J. G. (2001). *A Statistical History of the American Electorate*. CQ Press, Washington, DC.
WANG, W., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* **31** 980–991.
WRIGHT, F. A. and WRIGHT, A. A. (2018). How surprising was Trump's victory? Evaluations of the 2016 U.S. presidential election and a new poll aggregation model. *Elect. Stud.* **54** 81–89.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2078–2094
https://doi.org/10.1214/22-AOAS1708

LOG-GAUSSIAN COX PROCESS MODELING OF LARGE SPATIAL LIGHTNING DATA USING SPECTRAL AND LAPLACE APPROXIMATIONS

BY MEGAN L. GELSINGER[1,a], MARYCLARE GRIFFIN[2,d], DAVID MATTESON[1,b] AND JOSEPH GUINNESS[1,c]

[1]*Department of Statistics and Data Science, Cornell University,* [a]*mlg276@cornell.edu,* [b]*matteson@cornell.edu,* [c]*guinness@cornell.edu*

[2]*Department of Mathematics and Statistics, University of Massachusetts at Amherst,* [d]*maryclaregri@umass.edu*

Lightning is a destructive and highly visible product of severe storms, yet there is still much to be learned about the conditions under which lightning is most likely to occur. The GOES-16 and GOES-17 satellites, launched in 2016 and 2018 by NOAA and NASA, collect a wealth of data regarding individual lightning strike occurrence and potentially related atmospheric variables. The acute nature and inherent spatial correlation in lightning data renders standard regression analyses inappropriate. Further, computational considerations are foregrounded by the desire to analyze the immense and rapidly increasing volume of lightning data. We present a new computationally feasible method that combines spectral and Laplace approximations in an EM algorithm, denoted SLEM, to fit the widely popular log-Gaussian Cox process model to large spatial point pattern datasets. In simulations we find SLEM is competitive with contemporary techniques in terms of speed and accuracy. When applied to two lightning datasets, SLEM provides better out-of-sample prediction scores and quicker runtimes, suggesting its particular usefulness for analyzing lightning data which tend to have sparse signals.

REFERENCES

AICH, V., HOLZWORTH, R., GOODMAN, S., KULESHOV, Y., PRICE, C. and WILLIAMS, E. (2018). Lightning: A new essential climate variable. *Eos* **99**.

BACHL, F. E., LINDGREN, F., BORCHERS, D. L. and ILLIAN, J. B. (2019). Inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.* **10** 760–766.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton.

BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 823–841. MR1872069 https://doi.org/10.1111/1467-9868.00315

CLARK, S. K., WARD, D. S. and MAHOWALD, N. M. (2017). Parameterization-based uncertainty in future lightning flash density. *Geophys. Res. Lett.* **44** 2893–2901.

DIGGLE, P., ROWLINGSON, B. and SU, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics* **16** 423–434. MR2147534 https://doi.org/10.1002/env.712

DOC NOAA, N. and NASA (2017). GOES R series product definition and users' guide. *Atmos. Res.* **3**.

FINNEY, D. L., DOHERTY, R. M., WILD, O., STEVENSON, D. S., MACKENZIE, I. A. and BLYTH, A. M. (2018). A projected decrease in lightning under climate change. *Nat. Clim. Change* **8** 210–213.

GELFAND, A. E., KIM, H.-J., SIRMANS, C. F. and BANERJEE, S. (2003). Spatial modeling with spatially varying coefficient processes. *J. Amer. Statist. Assoc.* **98** 387–396. MR1995715 https://doi.org/10.1198/016214503000170

GELSINGER, M. L, GRIFFIN, M., MATTESON, D. and GUINNESS, J. (2023). Supplement to "Log-Gaussian cox process modeling of large spatial lightning data using spectral and laplace approximations." https://doi.org/10.1214/22-AOAS1708SUPP

GONÇALVES, F. B. and GAMERMAN, D. (2018). Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 157–175. MR3744716 https://doi.org/10.1111/rssb.12237

Key words and phrases. Spatial point pattern, Laplace approximation, expectation-maximization.

GOODMAN, S. J., BLAKESLEE, R. J., KOSHAK, W. J., MACH, D., BAILEY, J., BUECHLER, D., CAREY, L., SCHULTZ, C., BATEMAN, M. et al. (2013). The GOES-R geostationary lightning mapper (GLM). *Atmos. Res.* **125** 34–49.

GUAN, Y. and HARAN, M. (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. *J. Comput. Graph. Statist.* **27** 701–714. MR3890863 https://doi.org/10.1080/10618600.2018.1425625

GUAN, Y. and HARAN, M. (2020). Fast expectation-maximization algorithms for spatial generalized linear mixed models.

GUINNESS, J. and FUENTES, M. (2017). Circulant embedding of approximate covariances for inference from Gaussian data on large lattices. *J. Comput. Graph. Statist.* **26** 88–97. MR3610410 https://doi.org/10.1080/10618600.2016.1164534

HENDERSON, D. S., OTKIN, J. A. and MECIKALSKI, J. R. (2021). Evaluating convective initiation in high-resolution numerical weather prediction models using GOES-16 infrared brightness temperatures. *Mon. Weather Rev.* **149** 1153–1172.

HENDERSON, N. C. and VARADHAN, R. (2019). Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms. *J. Comput. Graph. Statist.* **28** 834–846. MR4045852 https://doi.org/10.1080/10618600.2019.1594835

HESTENES, M. R., STIEFEL, E. et al. (1952). *Methods of Conjugate Gradients for Solving Linear Systems* **49**. NBS, Washington, DC.

HUTCHINSON, M. F. (1989). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.* **18** 1059–1076. MR1031840 https://doi.org/10.1080/03610918908812806

ILLIAN, J. B., SØRBYE, S. H. and RUE, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Ann. Appl. Stat.* **6** 1499–1530. MR3058673 https://doi.org/10.1214/11-AOAS530

KATZFUSS, M., JUREK, M., ZILBER, D., GONG, W., GUINNESS, J., ZHANG, J. and SCHAEFER, F. (2021). GPvecchia: Scalable Gaussian-Process Computing. R package version 0.1.3.

KILINC, M. and BERINGER, J. (2007). The spatial and temporal distribution of lightning strikes and their relationship with vegetation type, elevation, and fire scars in the northern territory. *J. Climate* **20** 1161–1173.

KOTRONI, V. and LAGOUVARDOS, K. (2008). Lightning occurrence in relation with elevation, terrain slope, and vegetation cover in the Mediterranean. *J. Geophys. Res., Atmos.* **113**.

LEE, Y., KUMMEROW, C. D. and ZUPANSKI, M. (2021). A simplified method for the detection of convection using high-resolution imagery from GOES-16. *Atmos. Meas. Tech.* **14** 3755–3771.

LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

MARTINS, T. G., SIMPSON, D., LINDGREN, F. and RUE, H. (2013). Bayesian computing with INLA: New features. *Comput. Statist. Data Anal.* **67** 68–83. MR3079584 https://doi.org/10.1016/j.csda.2013.04.014

MECIKALSKI, J. R. and BEDKA, K. M. (2006). Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery. *Mon. Weather Rev.* **134** 49–78.

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019 https://doi.org/10.1111/1467-9469.00115

NOAA, N. G. D. C. Data Announcement 88-MGG-02, Digital relief of the Surface of the Earth.

PARK, J. and HARAN, M. (2021). Reduced-dimensional Monte Carlo maximum likelihood for latent Gaussian random field models. *J. Comput. Graph. Statist.* **30** 269–283. MR4270503 https://doi.org/10.1080/10618600.2020.1811106

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

SHIROTA, S. and GELFAND, A. E. (2016). Inference for log Gaussian Cox processes using an approximate marginal posterior. ArXiv preprint. Available at arXiv:1611.10359.

SYSTEM, G. C. O. (2016). The global observing system for climate: Implementation needs Technical Report No. 200 World Meteorological Organization Geneva, Switzerland.

TAYLOR, B. M. and DIGGLE, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *J. Stat. Comput. Simul.* **84** 2266–2284. MR3223624 https://doi.org/10.1080/00949655.2013.788653

TAYLOR, B. M., DAVIES, T. M., ROWLINGSON, B. S., DIGGLE, P. J. et al. (2013). lgcp: An R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *J. Stat. Softw.* **52** 1–40.

TAYLOR, B., DAVIES, T., ROWLINGSON, B. and DIGGLE, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *J. Stat. Softw.* **63** 1–48.

VARADHAN, R. and ROLAND, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* **35** 335–353. MR2418745 https://doi.org/10.1111/j.1467-9469.2007.00585.x

WOOD, A. T. A. and CHAN, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.* **3** 409–432. MR1323050 https://doi.org/10.2307/1390903

ZHOU, H., ALEXANDER, D. and LANGE, K. (2011). A quasi-Newton acceleration for high-dimensional optimization algorithms. *Stat. Comput.* **21** 261–273. MR2774856 https://doi.org/10.1007/s11222-009-9166-3

ZILBER, D. and KATZFUSS, M. (2021). Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. *Comput. Statist. Data Anal.* **153** 107081. MR4146817 https://doi.org/10.1016/j.csda.2020.107081

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2095–2117
https://doi.org/10.1214/22-AOAS1709

GRAPH-AWARE MODELING OF BRAIN CONNECTIVITY NETWORKS

BY YURA KIM[a], DANIEL KESSLER[b] AND ELIZAVETA LEVINA[c]

Department of Statistics, University of Michigan, [a]*kimyr@umich.edu,* [b]*kesslerd@umich.edu,* [c]*elevina@umich.edu*

Functional connections in the brain are frequently represented by weighted networks, with nodes representing locations in the brain and edges representing the strength of connectivity between these locations. One challenge in analyzing such data is that inference at the individual edge level is not particularly biologically meaningful; interpretation is more useful at the level of so-called functional systems or groups of nodes and connections between them; this is often called "graph-aware" inference in the neuroimaging literature. However, pooling over functional regions leads to significant loss of information and lower accuracy. Another challenge is correlation among edge weights within a subject which makes inference based on independence assumptions unreliable. We address both of these challenges with a linear mixed effects model, which accounts for functional systems and for edge dependence, while still modeling individual edge weights to avoid loss of information. The model allows for comparing two populations, such as patients and healthy controls, both at the functional regions level and at individual edge level, leading to biologically meaningful interpretations. We fit this model to resting state fMRI data on schizophrenic patients and healthy controls, obtaining interpretable results consistent with the schizophrenia literature.

REFERENCES

AINE, C. J., BOCKHOLT, H. J., BUSTILLO, J. R., CAÑIVE, J. M., CAPRIHAN, A., GASPAROVIC, C., HANLON, F. M., HOUCK, J. M., JUNG, R. E. et al. (2017). Multimodal neuroimaging in schizophrenia: Description and dissemination. *Neuroinformatics* **15** 343–364.

ANGRILLI, A., SPIRONELLI, C., ELBERT, T., CROW, T. J., MARANO, G. and STEGAGNO, L. (2009). Schizophrenia as failure of left hemispheric dominance for the phonological component of language. *PLoS ONE* **4** e4507. https://doi.org/10.1371/journal.pone.0004507

ARROYO RELIÓN, J. D., KESSLER, D., LEVINA, E. and TAYLOR, S. F. (2019). Network classification with applications to brain connectomics. *Ann. Appl. Stat.* **13** 1648–1677. MR4019153 https://doi.org/10.1214/19-AOAS1252

BAHRAMI, M., LAURIENTI, P. J. and SIMPSON, S. L. (2019). A Matlab toolbox for multivariate analysis of brain networks. *Hum. Brain Mapp.* **40** 175–186. https://doi.org/10.1002/hbm.24363

BAHRAMI, M., LAURIENTI, P. J., QUANDT, S. A., TALTON, J., POPE, C. N., SUMMERS, P., BURDETTE, J. H., CHEN, H., LIU, J. et al. (2017). The impacts of pesticide and nicotine exposures on functional brain networks in Latino immigrant workers. *NeuroToxicology* **62** 138–150.

BELILOVSKY, E., VAROQUAUX, G. and BLASCHKO, M. B. (2016). Testing for differences in Gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems* 29 (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 595–603. Curran Associates, Red Hook.

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

BULLMORE, E. T. (2012). Functional network endophenotypes of psychotic disorders. *Biol. Psychiatry* **71** 844–845.

BULLMORE, E. T. and BASSETT, D. S. (2011). Brain graphs: Graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* **7** 113–140. https://doi.org/10.1146/annurev-clinpsy-040510-143934

Key words and phrases. Neuroimaging, functional MRI, network analysis.

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

CHUNG, J., BRIDGEFORD, E., ARROYO, J., PEDIGO, B. D., SAAD-ELDIN, A., GOPALAKRISHNAN, V., XIANG, L., PRIEBE, C. E. and VOGELSTEIN, J. T. (2021). Statistical connectomics. *Annu. Rev. Stat. Appl.* **8** 463–492. MR4243556 https://doi.org/10.1146/annurev-statistics-042720-023234

CRADDOCK, R. C., HOLTZHEIMER, P. E., HU, X. P. and MAYBERG, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.* **62** 1619–1628.

CRADDOCK, R. C., JBABDI, S., YAN, C.-G., VOGELSTEIN, J. T., CASTELLANOS, F. X., MARTINO, A. D., KELLY, C., HEBERLEIN, K., COLCOMBE, S. et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* **10** 524–539. https://doi.org/10.1038/nmeth.2482

FRISTON, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2** 56–78.

FRISTON, K. J. and FRITH, C. D. (1995). Schizophrenia: A disconnection syndrome? *Clin. Neurosci.* **3** 89–97.

FIECAS, M., CRIBBEN, I., BAHKTIARI, R. and CUMMINE, J. (2017). A variance components model for statistical inference on functional connectivity networks. *NeuroImage* **149** 256–266. https://doi.org/10.1016/j.neuroimage.2017.01.051

FUJITA, A., TAKAHASHI, D. Y., BISOL BALARDIN, J., CALEBE VIDAL, M. and SATO, J. R. (2017). Correlation between graphs with an application to brain network analysis. *Comput. Statist. Data Anal.* **109** 76–92. MR3603642 https://doi.org/10.1016/j.csda.2016.11.016

HOCHBERG, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126 https://doi.org/10.1093/biomet/75.4.800

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

KIM, J., WOZNIAK, J. R., MUELLER, B. A., SHEN, X. and PAN, W. (2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage* **101** 681–694.

LANDIS, D., COURTNEY, W., DIERINGER, C., KELLY, R., KING, M., MILLER, B., WANG, R., WOOD, D., TURNER, J. A. et al. (2016). COINS data exchange: An open platform for compiling, curating, and disseminating neuroimaging data. *NeuroImage* **124** 1084–1088.

LI, J. (2015). *The Influence of Misspecification of Between-Subject and Within-Subject Covariance Structures in Hierarchical Growth Models*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Pittsburgh. MR3427391

LIANG, X., WANG, J., YAN, C., SHU, N., XU, K., GONG, G. and HE, Y. (2012). Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: A resting-state functional MRI study. *PLoS ONE* **7** e32766.

MITCHELL, R. L. C. and CROW, T. J. (2005). Right hemisphere language functions and schizophrenia: The forgotten hemisphere? *Brain* **128** 963–978. https://doi.org/10.1093/brain/awh466

NARAYAN, M., ALLEN, G. I. and TOMSON, S. (2015). Two sample inference for populations of graphical models with applications to functional connectivity. Preprint. Available at arXiv:1502.03853.

NARAYAN, M. and ALLEN, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Front. Neurosci.* **10** 108. https://doi.org/10.3389/fnins.2016.00108

PALANIYAPPAN, L., SIMMONITE, M., WHITE, T. P., LIDDLE, E. B. and LIDDLE, P. F. (2013). Neural primacy of the salience processing system in schizophrenia. *Neuron* **79** 814–828. https://doi.org/10.1016/j.neuron.2013.06.027

PAN, W., KIM, J., ZHANG, Y., SHEN, X. and WEI, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* **197** 1081–1095.

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

RIBOLSI, M., DASKALAKIS, Z. J., SIRACUSANO, A. and KOCH, G. (2014). Abnormal asymmetry of brain connectivity in schizophrenia. *Front. Human Neurosci.* **8** 1010. https://doi.org/10.3389/fnhum.2014.01010

SIMPSON, S. L., BAHRAMI, M. and LAURIENTI, P. J. (2019). A mixed-modeling framework for analyzing multitask whole-brain network data. *Netw. Neurosci.* **3** 307–324. https://doi.org/10.1162/netn_a_00065

SIMPSON, S. L. and LAURIENTI, P. J. (2015). A two-part mixed-effects modeling framework for analyzing whole-brain network data. *NeuroImage* **113** 310–319. https://doi.org/10.1016/j.neuroimage.2015.03.021

SMITH, S. M. (2012). The future of FMRI connectivity. *NeuroImage* **62** 1257–1266. https://doi.org/10.1016/j.neuroimage.2012.01.022

SMITH, S. M., MILLER, K. L., SALIMI-KHORSHIDI, G., WEBSTER, M., BECKMANN, C. F., NICHOLS, T. E., RAMSEY, J. D. and WOOLRICH, M. W. (2011). Network modelling methods for FMRI. *NeuroImage* **54** 875–891. https://doi.org/10.1016/j.neuroimage.2010.08.063

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.

SOBEL, M. E. and LINDQUIST, M. A. (2014). Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Amer. Statist. Assoc.* **109** 967–976. MR3265669 https://doi.org/10.1080/01621459.2014.922886

TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. MR3640191 https://doi.org/10.1080/10618600.2016.1193505

VAN DEN HEUVEL, M. P. and POL, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20** 519–534. https://doi.org/10.1016/j.euroneuro.2010.03.008

VAN DEN HEUVEL, M. P., MANDL, R. C. W., KAHN, R. S. and POL, H. E. H. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Hum. Brain Mapp.* **30** 3127–3141. https://doi.org/10.1002/hbm.20737

VAROQUAUX, G. and CRADDOCK, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage* **80** 405–415. https://doi.org/10.1016/j.neuroimage.2013.04.007

VENKATARAMAN, A., WHITFORD, T. J., WESTIN, C.-F., GOLLAND, P. and KUBICKI, M. (2012). Whole brain resting state functional connectivity abnormalities in schizophrenia. *Schizophr. Res.* **139** 7–12. https://doi.org/10.1016/j.schres.2012.04.021

WOOD, D., KING, M., LANDIS, D., COURTNEY, W., WANG, R., KELLY, R., TURNER, J. A. and CALHOUN, V. D. (2014). Harnessing modern web application technology to create intuitive and efficient data visualization and sharing tools. *Front. Neuroinform.* **8** 71.

XIA, M., WANG, J. and HE, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.

XIA, C. H., MA, Z., CUI, Z., BZDOK, D., THIRION, B., BASSETT, D. S., SATTERTHWAITE, T. D., SHINOHARA, R. T. and WITTEN, D. M. (2020). Multi-scale network regression for brain-phenotype associations. *Hum. Brain Mapp.*. https://doi.org/10.1002/hbm.24982

YEO, B. T. T., KRIENEN, F. M., SEPULCRE, J., SABUNCU, M. R., LASHKARI, D., HOLLINSHEAD, M., ROFFMAN, J. L., SMOLLER, J. W., ZÖLLEI, L. et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106** 1125–1165.

ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041

ZALESKY, A., COCCHI, L., FORNITO, A., MURRAY, M. M. and BULLMORE, E. (2012). Connectivity differences in brain networks. *NeuroImage* **60** 1055–1062. https://doi.org/10.1016/j.neuroimage.2012.01.068

ZHEN, Z., TIAN, J., QIN, W. and ZHANG, H. (2007). Partial correlation mapping of brain functional connectivity with resting state fMRI. *Proc. SPIE* **6511** 651112.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2118–2138
https://doi.org/10.1214/22-AOAS1710

BAYESIAN MODEL SELECTION: APPLICATION TO THE ADJUSTMENT OF FUNDAMENTAL PHYSICAL CONSTANTS

BY OLHA BODNAR[1,a] AND VIKTOR ERIKSSON[2,b]

[1]*Unit of Statistics, School of Business, Örebro University,* [a]*olha.bodnar@oru.se*
[2]*Department of Statistics, Uppsala University,* [b]*viktor.eriksson@statistik.uu.se*

A method originally suggested by Raymond Birge, using what came to be known as the *Birge ratio*, has been widely used in metrology and physics for the adjustment of fundamental physical constants, particularly in the periodic reevaluation carried out by the Task Group on Fundamental Physical Constants of CODATA (the Committee on Data of the International Science Council). The method involves increasing the reported uncertainties by a multiplicative factor large enough to make the measurement results mutually consistent. An alternative approach, predominant in the meta-analysis of medical studies, involves inflating the reported uncertainties by combining them, using the root sum of squares, with a sufficiently large constant (often dubbed *dark uncertainty*) that is estimated from the data.

In this contribution we establish a connection between the method based on the Birge ratio and the location-scale model, which allows one to combine the results of various studies, while the additive adjustment is reviewed in the usual context of random-effects models. Framing these alternative approaches as statistical models facilitates a quantitative comparison of them using statistical tools for model comparison. The intrinsic Bayes factor (IBF) is derived for the Berger and Bernardo reference prior, and then it is used to select a model for a set of measurements of the Newtonian constant of gravitation ("Big G") to estimate a consensus value for this constant and to evaluate the associated uncertainty. Our empirical findings support the method based on the Birge ratio. The same conclusion is reached when the IBF corresponding to the Jeffreys prior is used and also when the comparison is based on the Akaike information criterion (AIC). Finally, the results of a simulation study indicate that the suggested procedure for model selection provides clear guidance, even when the data comprise only a small number of measurements.

REFERENCES

ADES, A. E., LU, G. and HIGGINS, J. P. T. (2005). The interpretation of random-effects meta-analysis in decision models. *Med. Decis. Mak.* **25** 646–654. https://doi.org/10.1177/0272989X05282643

ALIGHANBARI, S., GIRI, G. S., CONSTANTIN, F. L., KOROBOV, V. I. and SCHILLER, S. (2020). Precise test of quantum electrodynamics and determination of fundamental constants with HD^+ ions. *Nature* **581** 152–158. https://doi.org/10.1038/s41586-020-2261-5

BERGER, J. O. and BERNARDO, J. M. (1992). On the development of reference priors. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) (J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, eds.) 35–60. Oxford Univ. Press, New York. MR1380269

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655 https://doi.org/10.1214/07-AOS587

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065 https://doi.org/10.2307/2291387

BERGER, J. O. and PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (P. Lahiri, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 135–207. IMS, Beachwood, OH. MR2000753 https://doi.org/10.1214/lnms/1215540968

BIRGE, R. T. (1932). The calculation of errors by the method of the least squares. *Phys. Rev.* **40** 207–227.

Key words and phrases. Intrinsic Bayes factor, Birge ratio method, location-scale model, random-effects model, reference prior, meta-analysis, interlaboratory comparison study, Newtonian constant of gravitation.

BODNAR, O. and ELSTER, C. (2014a). Analytical derivation of the reference prior by sequential maximization of Shannon's mutual information in the multi-group parameter case. *J. Statist. Plann. Inference* **147** 106–116. MR3151849 https://doi.org/10.1016/j.jspi.2013.11.003

BODNAR, O. and ELSTER, C. (2014b). On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51** 516.

BODNAR, O. and ELSTER, C. (2021). Assessing laboratory effects in key comparisons with two transfer standards measured in two petals: A Bayesian approach. In *Frontiers in Statistical Quality Control* 13. *Front. Stat. Qual. Control* 359–376. Springer, Cham. MR4376017 https://doi.org/10.1007/978-3-030-67856-2_20

BODNAR, O. and ERIKSSON, V. (2023a). Supplement A to "Bayesian model selection: Application to the adjustment of fundamental physical constants." https://doi.org/10.1214/22-AOAS1710SUPPA

BODNAR, O. and ERIKSSON, V. (2023b). Supplement B to "Bayesian model selection: Application to the adjustment of fundamental physical constants." https://doi.org/10.1214/22-AOAS1710SUPPB

BODNAR, O. and ERIKSSON, V. (2023c). Supplement C to "Bayesian model selection: Application to the adjustment of fundamental physical constants." https://doi.org/10.1214/22-AOAS1710SUPPC

BODNAR, O. and ERIKSSON, V. (2023d). Supplement D to "Bayesian model selection: Application to the adjustment of fundamental physical constants." https://doi.org/10.1214/22-AOAS1710SUPPD

BODNAR, O., LINK, A. and ELSTER, C. (2016). Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Anal.* **11** 25–45. MR3447090 https://doi.org/10.1214/14-BA933

BODNAR, O., MUHUMUZA, R. N. and POSSOLO, A. (2020). Bayesian inference for heterogeneity in meta-analysis. *Metrologia* **57** 064004.

BODNAR, O., ELSTER, C., FISCHER, J., POSSOLO, A. and TOMAN, B. (2016). Evaluation of uncertainty in the adjustment of fundamental constants. *Metrologia* **53** S46.

BODNAR, O., LINK, A., ARENDACKÁ, B., POSSOLO, A. and ELSTER, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Stat. Med.* **36** 378–399. MR3582981 https://doi.org/10.1002/sim.7156

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

BROCKWELL, S. E. and GORDON, I. R. (2001). A comparison of statistical methods for meta-analysis. *Stat. Med.* **20** 825–840. https://doi.org/10.1002/sim.650

CLARKE, B. and YUAN, A. (2004). Partial information reference priors: Derivation and interpretations. *J. Statist. Plann. Inference* **123** 313–345. MR2062985 https://doi.org/10.1016/S0378-3758(03)00157-5

FERNÁNDEZ, C. and STEEL, M. F. J. (1999). Reference priors for the general location-scale model. *Statist. Probab. Lett.* **43** 377–384. MR1707947 https://doi.org/10.1016/S0167-7152(98)00276-4

GENEST, C. and SCHERVISH, M. J. (1985). Modeling expert judgments for Bayesian updating. *Ann. Statist.* **13** 1198–1212. MR0803766 https://doi.org/10.1214/aos/1176349664

GIVENS, G. H. and HOETING, J. A. (2013). *Computational Statistics*, 2nd ed. *Wiley Series in Computational Statistics*. Wiley, Hoboken, NJ. MR3236433

PARTICLE DATA GROUP, ZYLA, P. A. et al. (2020). Review of particle physics. *Prog. Theor. Exp. Phys.* **2020** 083C01. https://doi.org/10.1093/ptep/ptaa104

GUOLO, A. and VARIN, C. (2017). Random-effects meta-analysis: The number of studies matters. *Stat. Methods Med. Res.* **26** 1500–1518. MR3661007 https://doi.org/10.1177/0962280215583568

HANNIG, J., FENG, Q., IYER, H., WANG, C. M. and LIU, X. (2018). Fusion learning for inter-laboratory comparisons. *J. Statist. Plann. Inference* **195** 64–79. MR3760838 https://doi.org/10.1016/j.jspi.2017.09.011

HARDY, R. J. and THOMPSON, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Stat. Med.* **17** 841–856. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<841::aid-sim781>3.0.co;2-d

HELD, L. and SABANÉS BOVÉ, D. (2014). *Applied Statistical Inference*: *Likelihood and Bayes*. Springer, Heidelberg. MR3155114 https://doi.org/10.1007/978-3-642-37887-4

HIGGINS, J. P. T., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. *J. Roy. Statist. Soc. Ser. A* **172** 137–159. MR2655609 https://doi.org/10.1111/j.1467-985X.2008.00552.x

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A* **186** 453–461. MR0017504 https://doi.org/10.1098/rspa.1946.0056

JONES, H. E., ADES, A. E., SUTTON, A. J. and WELTON, N. J. (2018). Use of a random effects meta-analysis in the design and analysis of a new clinical trial. *Stat. Med.* **37** 4665–4679. MR3883432 https://doi.org/10.1002/sim.7948

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

KOEPKE, A., LAFARGE, T., POSSOLO, A. and TOMAN, B. (2017). Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia* **54** S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e

LAMBERT, P. C., SUTTON, A. J., BURTON, P. R., ABRAMS, K. R. and JONES, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat. Med.* **24** 2401–2428. MR2151713 https://doi.org/10.1002/sim.2112

MANDEL, J. and PAULE, R. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. *Anal. Chem.* **42** 1194–1197. https://doi.org/10.1021/ac60293a019

MEIJA, J. and POSSOLO, A. (2017). Data reduction framework for standard atomic weights and isotopic compositions of the elements. *Metrologia* **54** 229–238.

MERKATAS, C., TOMAN, B., POSSOLO, A. and SCHLAMMINGER, S. (2019). Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation. *Metrologia* **56** 054001. https://doi.org/10.1088/1681-7575/ab3365

MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. *Rev. Modern Phys.* **88** 035009.

NEWELL, D. B., CABIATI, F., FISCHER, J., FUJII, K., KARSHENBOIM, S. G., MARGOLIS, H. S., DE MIRANDES, E., MOHR, P. J., NEZ, F. et al. (2018). The CODATA 2017 values of h, e, k, and N_A for the revision of the SI. *Metrologia* **55** L13–L16.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. MR1325379

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York. https://doi.org/10.1007/b98882

POSSOLO, A., KOEPKE, A., NEWTON, D. and WINCHESTER, M. R. (2021). Decision tree for key comparisons. *J. Res. Natl. Inst. Stand. Technol.* **126** 126007. https://doi.org/10.6028/jres.126.007

ROTHLEITNER, C. and SCHLAMMINGER, S. (2017). Invited Review Article: Measurements of the Newtonian constant of gravitation, *G*. *Rev. Sci. Instrum.* **88** 111101. https://doi.org/10.1063/1.4994619

RUKHIN, A. L. (2003). Two procedures of meta-analysis in clinical trials and interlaboratory studies. *Tatra Mt. Math. Publ.* **26** 155–168. MR2055171

RUKHIN, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 451–469. MR3065475 https://doi.org/10.1111/j.1467-9868.2012.01047.x

THOMPSON, M. and ELLISON, S. L. (2011). Dark uncertainty. *Accredit. Qual. Assur.* **16** 483–487. https://doi.org/10.1007/s00769-011-0803-0

TIESINGA, E., MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2021). CODATA recommended values of the fundamental physical constants: 2018. *Rev. Modern Phys.* **93** 025010.

TOMAN, B., FISCHER, J. and ELSTER, C. (2012). Alternative analyses of measurements of the Planck constant. *Metrologia* **49** 567–571.

TOMAN, B. and POSSOLO, A. (2009). Laboratory effects models for interlaboratory comparisons. *Accredit. Qual. Assur.* **14** 553–563. https://doi.org/10.1007/s00769-009-0547-2

TOMAN, B. and POSSOLO, A. (2010). Erratum to: Laboratory effects models for interlaboratory comparisons. *Accredit. Qual. Assur.* **15** 653–654. https://doi.org/10.1007/s00769-010-0707-4

TURNER, R. M., JACKSON, D., WEI, Y., THOMPSON, S. G. and HIGGINS, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat. Med.* **34** 984–998. MR3310676 https://doi.org/10.1002/sim.6381

VERONIKI, A. A., JACKSON, D., BENDER, R., KUSS, O., LANGAN, D., HIGGINS, J. P. T., KNAPP, G. and SALANTI, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Res. Synth. Methods* **10** 23–43. https://doi.org/10.1002/jrsm.1319

WEISE, K. and WÖGER, W. (2000). Removing model and data non-conformity in measurement evaluation. *Meas. Sci. Technol.* **11** 1649.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2139–2164
https://doi.org/10.1214/22-AOAS1712

LEVERAGING POPULATION OUTCOMES TO IMPROVE THE GENERALIZATION OF EXPERIMENTAL RESULTS: APPLICATION TO THE JTPA STUDY

BY MELODY HUANG[1,a], NAOKI EGAMI[2,b], ERIN HARTMAN[3,c] AND LUKE MIRATRIX[4,d]

[1]*Department of Statistics, University of California, Berkeley,* [a]*melodyyhuang@berkeley.edu*
[2]*Department of Political Science, Columbia University,* [b]*naoki.egami@columbia.edu*
[3]*Departments of Political Science and Statistics, University of California, Berkeley,* [c]*ekhartman@berkeley.edu*
[4]*Graduate School of Education, Harvard University,* [d]*lmiratrix@g.harvard.edu*

Generalizing causal estimates in randomized experiments to a broader target population is essential for guiding decisions by policymakers and practitioners in the social and biomedical sciences. While recent papers have developed various weighting estimators for the population average treatment effect (PATE), many of these methods result in large variance because the experimental sample often differs substantially from the target population and estimated sampling weights are extreme. We investigate this practical problem motivated by an evaluation study of the Job Training Partnership Act (JTPA), where we examine how well we can generalize the causal effect of job training programs beyond a specific population of economically disadvantaged adults and youths. In particular, we propose post-residualized weighting in which we use the outcome measured in the observational population data to build a flexible predictive model (e.g., with machine learning) and residualize the outcome in the experimental data before using conventional weighting methods. We show that the proposed PATE estimator is consistent under the same assumptions required for existing weighting methods, importantly without assuming the correct specification of the predictive model. We demonstrate the efficiency gains from this approach through our JTPA application: we find a reduction of between 5% and 25% in variance.

REFERENCES

ATHEY, S., CHETTY, R. and IMBENS, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676.

ATHEY, S., CHETTY, R., IMBENS, G. W. and KANG, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report, National Bureau of Economic Research.

BALDASSARRI, D. and ABASCAL, M. (2017). Field experiments across the social sciences. *Annu. Rev. Sociol.* **43** 41–73.

BANERJEE, A. V. and DUFLO, E. (2009). The experimental approach to development economics. *Ann. Rev. Econ.* **1** 151–178.

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA* **113** 7345–7352.

BLOOM, H. S. ORR, L. CAVE, G. BELL, S. and DOOLITTLE, F. (1993). The national JTPA study. Title II-a impacts on earnings and employment at 18 months. Bethesda, MD: Abt Associates.

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

BUCHANAN, A. L., HUDGENS, M. G., COLE, S. R., MOLLAN, K. R., SAX, P. E., DAAR, E. S., ADIMORA, A. A., ERON, J. J. and MUGAVERO, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. Roy. Statist. Soc. Ser. A* **181** 1193–1209. MR3876388 https://doi.org/10.1111/rssa.12357

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Technical Report, National Bureau of Economic Research.

Key words and phrases. Generalizability, external validity, randomized experiments, causal inference.

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115. https://doi.org/10.1093/aje/kwq084

COLNET, B., MAYER, I., CHEN, G., DIENG, A., LI, R., VAROQUAUX, G., VERT, J.-P., JOSSE, J. and YANG, S. (2020). Causal inference methods for combining randomized trials and observational studies: A review. arXiv preprint arXiv:2011.08047.

DAHABREH, I. J., ROBERTSON, S. E., TCHETGEN, E. J., STUART, E. A. and HERNÁN, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694. MR3999190 https://doi.org/10.1111/biom.13009

DEATON, A. and CARTWRIGHT, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210** 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

DING, P. (2021). Two seemingly paradoxical results in linear models: The variance inflation factor and the analysis of covariance. *J. Causal Inference* **9** 1–8. MR4289523 https://doi.org/10.1515/jci-2019-0023

EGAMI, N. and HARTMAN, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *J. Roy. Statist. Soc. Ser. A* **184** 1524–1548. MR4344647 https://doi.org/10.1111/rssa.12734

EGAMI, N. and HARTMAN, E. (2022). Elements of external validity: Framework, design, and analysis *Amer. Polit. Sci. Rev.* **First View**, 1–19. https://doi.org/10.1017/S0003055422000880

FALK, A. and HECKMAN, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science* **326** 535–538.

FORD, I. and NORRIE, J. (2016). Pragmatic trials. *N. Engl. J. Med.* **375** 454–463. PMID: 27518663. https://doi.org/10.1056/NEJMra1510059

FREEDMAN, D. A. (2008). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196. MR2415599 https://doi.org/10.1214/07-AOAS143

GERBER, A. S. and GREEN, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *Amer. Polit. Sci. Rev.* **94** 653–663.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 25–46.

HARTMAN, E., GRIEVE, R., RAMSAHAI, R. and SEKHON, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *J. Roy. Statist. Soc. Ser. A* **178** 757–778. MR3348358 https://doi.org/10.1111/rssa.12094

HUANG, M., EGAMI, N., HARTMAN, E. and MIRATRIX, L. (2023). Supplement to "Leveraging population outcomes to improve the generalization of experimental results: Application to the JTPA study." https://doi.org/10.1214/22-AOAS1712SUPP

HUBER, J. (2013). Is theory getting lost in the "identification revolution"? The Monkey Cage.

JACOB, D. (2020). Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. arXiv preprint arXiv:2007.02852.

KALLUS, N. and MAO, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv:2003.12408.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KERN, H. L., STUART, E. A., HILL, J. and GREEN, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* **9** 103–127. https://doi.org/10.1080/19345747.2015.1060282

LALONDE, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econ. Rev.* 604–620. https://doi.org/10.2307/1806062

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

MIRATRIX, L. W., SEKHON, J. S., THEODORIDIS, A. G. and CAMPOS, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Polit. Anal.* **26** 275–291. https://doi.org/10.1017/pan.2018.1

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statist. Sci.* **5** 465–472. MR1092986

NGUYEN, T. Q., EBNESAJJAD, C., COLE, S. R. and STUART, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* **11** 225–247. MR3634322 https://doi.org/10.1214/16-AOAS1001

O'MUIRCHEARTAIGH, C. and HEDGES, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 195–210. MR3234340 https://doi.org/10.1111/rssc.12037

PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595. MR3300360 https://doi.org/10.1214/14-STS486

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

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. With comments and a rejoinder by the author. MR1962487 https://doi.org/10.1214/ss/1042727942

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688. https://doi.org/10.1037/h0037350

RUBIN, D. B. (1980). Discussion of 'Randomization analysis of experimental data: The Fisher randomization test comment' by Basu. *J. Amer. Statist. Assoc.* **75** 591–593. https://doi.org/10.2307/2287653

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SALES, A. C., HANSEN, B. B. and ROWAN, B. (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *J. Educ. Behav. Stat.* **43** 3–31. https://doi.org/10.3102/1076998617731518

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

STUART, E. A. and RHODES, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Eval. Rev.* **41** 357–388. https://doi.org/10.1177/0193841X16660663

STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386. MR2898850 https://doi.org/10.1111/j.1467-985X.2010.00673.x

TIPTON, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

WESTREICH, D. and COLE, S. R. (2010). Invited commentary: Positivity in practice. *Amer. J. Epidemiol.* **171** 674–677. https://doi.org/10.1093/aje/kwp436

WORD, E. R. et al. (1990). The state of Tennessee's student/teacher achievement ratio (STAR) project: Technical report (1985–1990). Final summary report. Nashville: Tennessee Dept. Education.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2165–2191
https://doi.org/10.1214/22-AOAS1713

ESTIMATING CAUSAL EFFECTS OF HIV PREVENTION INTERVENTIONS WITH INTERFERENCE IN NETWORK-BASED STUDIES AMONG PEOPLE WHO INJECT DRUGS

BY TINGFANG LEE[1,a], ASHLEY L. BUCHANAN[1,b], NATALLIA V. KATENKA[2,c], LAURA FORASTIERE[3,d], M. ELIZABETH HALLORAN[4,5,e], SAMUEL R. FRIEDMAN[6,f] AND GEORGIOS NIKOLOPOULOS[7,g]

[1]*Department of Pharmacy Practice, University of Rhode Island,* [a]*tingfanglee@uri.edu,* [b]*buchanan@uri.edu*
[2]*Department of Computer Science and Statistics, University of Rhode Island,* [c]*nkatenka@uri.edu*
[3]*Department of Biostatistics, Yale School of Public Health,* [d]*laura.forastiere@yale.edu*
[4]*Biostatistics, Bioinformatics, and Epidemiology Program, Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center*
[5]*Department of Biostatistics, University of Washington,* [e]*betz@fredhutch.org*
[6]*Department of Population Health, NYU Grossman School of Medicine,* [f]*Samuel.Friedman@nyulangone.org*
[7]*Medical School, University of Cyprus,* [g]*nikolopoulos.georgios@ucy.ac.cy*

Evaluating causal effects in the presence of interference is challenging in network-based studies of hard-to-reach populations. Like many such populations, people who inject drugs (PWID) are embedded in social networks and one person's treatment can affect the outcomes of others in the network. In our setting, the study design is observational with a nonrandomized network-based HIV prevention intervention. Information is available on each participant and their connections that confer possible HIV risk through injection and sexual behaviors. We considered two inverse probability weighted (IPW) estimators to quantify the population-level spillover effects of nonrandomized interventions on subsequent health outcomes. We demonstrated that these two IPW estimators are consistent, asymptotically normal, and derived a closed-form estimator for the asymptotic variance, while allowing for overlapping interference sets (groups of individuals in which the interference is assumed possible). A simulation study was conducted to evaluate the finite-sample performance of the estimators. We analyzed data from the Transmission Reduction Intervention Project which ascertained a network of PWID and their contacts in Athens, Greece, from 2013 to 2015. We evaluated the effects of community alerts on subsequent HIV risk behavior in this observed network, where the connections or links between participants were defined by using substances or having unprotected sex together. In the study, community alerts were distributed to inform people of recent HIV infections among individuals in close proximity in the observed network. The estimates of the risk differences for spillover, using either IPW estimator demonstrated a protective effect. The results suggest that HIV risk behavior could be mitigated by exposure to a community alert when an increased risk of HIV is detected in the network.

REFERENCES

ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* **11** 1912–1947. MR3743283 https://doi.org/10.1214/16-AOAS1005

BASSE, G. and FELLER, A. (2018). Analyzing two-stage experiments in the presence of interference. *J. Amer. Statist. Assoc.* **113** 41–55. MR3803438 https://doi.org/10.1080/01621459.2017.1323641

Key words and phrases. Causal interference, interference/dissemination, network studies, people who use drugs, HIV/AIDS, inverse probability weights.

BENJAMIN-CHUNG, J., ARNOLD, B., BERGER, D., LUBY, S., MIGUEL, E., COLFORD, J. and HUBBARD, A. (2017). Spillover effects in epidemiology: Parameters, study designs and methodological considerations. *Int. J. Epidemiol.* **47** 332–347.

BOOS, D. and STEFANSKI, L. (2013). M-Estimation (Estimating Equations). 297–337.

BUCHANAN, A. L., VERMUND, S. H., FRIEDMAN, S. R. and SPIEGELMAN, D. (2018). Assessing individual and disseminated effects in network-randomized studies. *Amer. J. Epidemiol.* **187** 2449–2459. https://doi.org/10.1093/aje/kwy149

CLAUSET, A., NEWMAN, M. E. J. and MOORE, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* **70** 066111.

DESROSIERS, A., KUMAR, P., DAYAL, A., ALEX, L., AKRAM, A. and BETANCOURT, T. (2020). Diffusion and spillover effects of an evidence-based mental health intervention among peers and caregivers of high risk youth in Sierra Leone: Study protocol. *BMC Psychiatry* **20**.

FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *J. Amer. Statist. Assoc.* **116** 901–918. MR4270033 https://doi.org/10.1080/01621459.2020.1768100

FRIEDMAN, S. and ARAL, S. (2001). Social networks, risk-potential networks, health, and disease. *J. Urban Health* **73** 411–8.

FRIEDMAN, S. R., DOWNING, M. J., SMYRNOV, P., NIKOLOPOULOS, G., SCHNEIDER, J. A., LIVAK, B., MAGIORKINIS, G., SLOBODIANYK, L., VASYLYEVA, T. I. et al. (2014). Socially-integrated transdisciplinary HIV prevention. *AIDS Behav.* **18** 1821–1834.

GHOSH, D., KRISHNAN, A., GIBSON, B., BROWN, S.-E., LATKIN, C. A. and ALTICE, F. L. (2017). Social network strategies to address HIV prevention and treatment continuum of care among at-risk and HIV-infected substance users: A systematic scoping review. *AIDS Behav.* **21** 1183–1207.

GIALLOUROS, G., PANTAVOU, K., PAMPAKA, D., PAVLITINA, E., PIOVANI, D., BONOVAS, S. and NIKOLOPOULOS, G. K. (2021). Drug injection-related and sexual behavior changes in drug injecting networks after the transmission reduction intervention project (TRIP): A social network-based study in Athens, Greece. *Int. J. Environ. Res. Public Health* **18** 2388.

HADJIKOU, A., PAVLOPOULOU, I. D., PANTAVOU, K., GEORGIOU, A., WILLIAMS, L. D., CHRISTAKI, E., VOSKARIDES, K., LAVRANOS, G., LAMNISOS, D. et al. (2021). Drug injection-related norms and high-risk behaviors of people who inject drugs in Athens, Greece. *AIDS Res. Hum. Retrovir.* **37** 130–138.

HAYES, R. J., ALEXANDER, N. D., BENNETT, S. and COUSENS, S. N. (2000). Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Stat. Methods Med. Res.* **9** 95–116. https://doi.org/10.1177/096228020000900203

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

LANSKY, A., FINLAYSON, T., JOHNSON, C., HOLTZMAN, D., WEJNERT, C., MITSCH, A., GUST, D., CHEN, R., MIZUNO, Y. et al. (2014). Estimating the number of persons who inject drugs in the United States by meta-analysis to calculate national rates of HIV and hepatitis C virus infections. *PLoS ONE* **9** e97596.

LEE, T., BUCHANAN, A., KATENKA, N., FORASTIERE, L., HALLORAN, M., FRIEDMAN, S. and NIKOLOPOULOS, G. (2023). Appendices: Estimating causal effects of non-randomized HIV prevention interventions with interference in network-based studies among people who inject drugs." https://doi.org/10.1214/22-AOAS1713SUPPA, https://doi.org/10.1214/22-AOAS1713SUPPB

LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564 https://doi.org/10.1080/01621459.2013.844698

LIU, L., HUDGENS, M. G. and BECKER-DREPS, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika* **103** 829–842. MR3620442 https://doi.org/10.1093/biomet/asw047

MATHERS, B. M., DEGENHARDT, L., PHILLIPS, B., WIESSING, L., HICKMAN, M., STRATHDEE, S. A., WODAK, A., PANDA, S., TYNDALL, M. et al. (2008). Global epidemiology of injecting drug use and HIV among people who inject drugs: A systematic review. *Lancet* **372** 1733–1745.

NIKOLOPOULOS, G., SYPSA, V., BONOVAS, S., PARASKEVIS, D., MALLIORI, M., HATZAKIS, A. and FRIEDMAN, S. (2015). Big events in Greece and HIV infection among people who inject drugs. *Subst. Use Misuse* **50** 1–14.

NIKOLOPOULOS, G. K., PAVLITINA, E., MUTH, S. Q., SCHNEIDER, J., PSICHOGIOU, M., WILLIAMS, L. D., PARASKEVIS, D., SYPSA, V., MAGIORKINIS, G. et al. (2016). A network intervention that locates and intervenes with recently HIV-infected persons: The Transmission Reduction Intervention Project (TRIP). *Sci. Rep.* **6**.

OGBURN, E. L. and VANDERWEELE, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29** 559–578. MR3300359 https://doi.org/10.1214/14-STS501

PAMPAKA, D., PANTAVOU, K., GIALLOUROS, G., PAVLITINA, E., WILLIAMS, L. D., PIOVANI, D., BONOVAS, S. and NIKOLOPOULOS, G. K. (2021). Mental health and perceived access to care among people who inject drugs in Athens, Greece. *J. Clin. Med.* **10**. https://doi.org/10.3390/jcm10061181

PARASKEVIS, D., NIKOLOPOULOS, G., FOTIOU, A., TSIARA, C., PARASKEVA, D., SYPSA, V., LAZANAS, M. K., GARGALIANOS-KAKOLYRIS, P., PSICHOGIOU, M. et al. (2013). Economic recession and emergence of an HIV-1 outbreak among drug injectors in Athens metropolitan area: A longitudinal study. *PLoS ONE* **8** e78941.

PREJEAN, J., SONG, R., HERNANDEZ, A., ZIEBELL, R., GREEN, T., WALKER, F., LIN, L. S., AN, Q., MERMIN, J. et al. (2011). Estimated HIV incidence in the United States, 2006–2009. *PLoS ONE* **6** e17502.

PSICHOGIOU, M., GIALLOUROS, G., PANTAVOU, K., PAVLITINA, E., PAPADOPOULOU, M., WILLIAMS, L. D., HADJIKOU, A., KAKALOU, E., SKOUTELIS, A. et al. (2019). Identifying, linking, and treating people who inject drugs and were recently infected with HIV in the context of a network-based intervention. *AIDS Care*.

REWLEY, J., FAWZI, M. C. S., MCADAM, K., KAAYA, S., LIU, Y., TODD, J., ANDREW, I. and ONNELA, J. P. (2020). Evaluating spillover of HIV knowledge from study participants to their network members in a stepped-wedge behavioural intervention in Tanzania. *BMJ Open* **10**.

RUBIN, D. B. (1980). Discussion of randomization analysis of experimental data in the Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

SYPSA, V., PARASKEVIS, D., MALLIORI, M., NIKOLOPOULOS, G., PANOPOULOS, A., KANTZANOU, M., KATSOULIDOU, A., PSICHOGIOU, M., FOTIOU, A. et al. (2014). Homelessness and other risk factors for HIV infection in the current outbreak among injection drug users in Athens, Greece. *Amer. J. Publ. Health* **105**.

TCHETGEN, E. J. and COULL, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika* **93** 1003–1010. MR2285086 https://doi.org/10.1093/biomet/93.4.1003

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J. and HALLORAN, M. E. (2014). Interference and sensitivity analysis. *Statist. Sci.* **29** 687–706. MR3300366 https://doi.org/10.1214/14-STS479

WILLIAMS, L. D., KOSTAKI, E.-G., PAVLITINA, E., PARASKEVIS, D., HATZAKIS, A., SCHNEIDER, J., SMYRNOV, P., HADJIKOU, A., NIKOLOPOULOS, G. K. et al. (2018). Pockets of HIV non-infection within highly-infected risk networks in Athens, Greece. *Front. Microbiol.* **9** 1825. https://doi.org/10.3389/fmicb.2018.01825

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2192–2211
https://doi.org/10.1214/22-AOAS1714

USING PERSISTENT HOMOLOGY TOPOLOGICAL FEATURES TO CHARACTERIZE MEDICAL IMAGES: CASE STUDIES ON LUNG AND BRAIN CANCERS

BY CHUL MOON[1,a], QIWEI LI[2,b] AND GUANGHUA XIAO[3,c]

[1]*Department of Statistical Science, Southern Methodist University,* [a]*chulm@smu.edu*
[2]*Department of Mathematical Sciences, University of Texas at Dallas,* [b]*qiwei.li@utdallas.edu*
[3]*Quantitative Biomedical Research Center, Department of Population & Data Sciences and Department of Bioinformatics, University of Texas Southwestern Medical Center,* [c]*guanghua.xiao@utsouthwestern.edu*

Tumor shape is a key factor that affects tumor growth and metastasis. This paper proposes a topological feature computed by persistent homology to characterize tumor progression from digital pathology and radiology images and examines its effect on the time-to-event data. The proposed topological features are invariant to scale-preserving transformation and can summarize various tumor shape patterns. The topological features are represented in functional space and used as functional predictors in a functional Cox proportional hazards model. The proposed model enables interpretable inference about the association between topological shape features and survival risks. Two case studies are conducted using consecutive 133 lung cancer and 77 brain tumor patients. The results of both studies show that the topological features predict survival prognosis after adjusting clinical variables, and the predicted high-risk groups have worse survival outcomes than the low-risk groups. Also, the topological shape features found to be positively associated with survival hazards are irregular and heterogeneous shape patterns which are known to be related to tumor progression.

REFERENCES

ADAMS, H., EMERSON, T., KIRBY, M., NEVILLE, R., PETERSON, C., SHIPMAN, P., CHEPUSHTANOVA, S., HANSON, E., MOTTA, F. et al. (2017). Persistence images: A stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18** Paper No. 8. MR3625712

BENDER, R., AUGUSTIN, T. and BLETTNER, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24** 1713–1723. MR2137646 https://doi.org/10.1002/sim.2059

BERRY, E., CHEN, Y.-C., CISEWSKI-KEHE, J. and FASY, B. T. (2020). Functional summaries of persistence diagrams. *J. Appl. Comput. Topol.* **4** 211–262. MR4096338 https://doi.org/10.1007/s41468-020-00048-w

BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2018). Radiologic image-based statistical shape analysis of brain tumours. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1357–1378. MR3873711 https://doi.org/10.1111/rssc.12272

BIANCONI, F., FRAVOLINI, M. L., BELLO-CEREZO, R., MINESTRINI, M., SCIALPI, M. and PALUMBO, B. (2018). Evaluation of shape and textural features from CT as prognostic biomarkers in non-small cell lung cancer. *Anticancer Res.* **38** 2155–2160. https://doi.org/10.21873/anticanres.12456

BONDY, M. L., SCHEURER, M. E., MALMER, B., BARNHOLTZ-SLOAN, J. S., DAVIS, F. G., IL'YASOVA, D., KRUCHKO, C., MCCARTHY, B. J., RAJARAMAN, P. et al. (2008). Brain tumor epidemiology: Consensus from the brain tumor epidemiology consortium. *Cancer* **113** 1953–1968.

BOOKSTEIN, F. L. (1997). *Morphometric Tools for Landmark Data*: *Geometry and Biology*. Cambridge Univ. Press, Cambridge. MR1469220

BRÚ, A., CASERO, D., DE FRANCISCIS, S. and HERRERO, M. A. (2008). Fractal analysis and tumour growth. *Math. Comput. Modelling* **47** 546–559. MR2396790 https://doi.org/10.1016/j.mcm.2007.02.033

BUBENIK, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

Key words and phrases. Topological data analysis, tumor shape, functional data analysis, survival analysis, Cox proportional hazards model.

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 255–308. MR2476414 https://doi.org/10.1090/S0273-0979-09-01249-X

CHATZISTAMOU, I., RODRIGUEZ, J., JOUFFROY, T., GIROD, A., POINT, D., SKLAVOUNOU, A., KITTAS, C., SASTRE-GARAU, X. and KLIJANIENKO, J. (2010). Prognostic significance of tumor shape and stromal chronic inflammatory infiltration in squamous cell carcinomas of the oral tongue. *Journal of Oral Pathology & Medicine* **39** 667–671.

CHAZAL, F. and MICHEL, B. (2021). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Front. Artif. Intell.* **4** 667963. https://doi.org/10.3389/frai.2021.667963

CHAZAL, F., FASY, B., LECCI, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2017). Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* **18** Paper No. 159. MR3813808

CHEN, A. X. and RABADÁN, R. (2017). A fast semi-automatic segmentation tool for processing brain tumor images. In *Towards Integrative Machine Learning and Knowledge Extraction* 170–181. Springer, Berlin.

CHEN, K., CHEN, K., MÜLLER, H.-G. and WANG, J.-L. (2011). Stringing high-dimensional data for functional analysis. *J. Amer. Statist. Assoc.* **106** 275–284. MR2816720 https://doi.org/10.1198/jasa.2011.tm10314

CHEN, Y.-C., WANG, D., RINALDO, A. and WASSERMAN, L. (2015). Statistical analysis of persistence intensity functions. arXiv e-prints.

COUPRIE, M., BEZERRA, F.-N. and BERTRAND, G. (2001). Topological operators for grayscale image processing. *Journal of Electronic Imaging* **10** 1003–1015.

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

CRAWFORD, L., MONOD, A., CHEN, A. X., MUKHERJEE, S. and RABADÁN, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Amer. Statist. Assoc.* **115** 1139–1150. MR4143455 https://doi.org/10.1080/01621459.2019.1671198

DLOTKO, P. (2015). Cubical complex. In *GUDHI User and Reference Manual* GUDHI Editorial Board.

EFRON, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565. MR0451514

EISENHAUER, E. A., THERASSE, P., BOGAERTS, J., SCHWARTZ, L. H., SARGENT, D., FORD, R., DANCEY, J., ARBUCK, S., GWYTHER, S. et al. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45** 228–247.

FABBRI, R., COSTA, L. D. F., TORELLI, J. C. and BRUNO, O. M. (2008). 2D Euclidean distance transform algorithms: A comparative survey. *ACM Comput. Surv.* **40** 1–44.

GELLAR, J. E., COLANTUONI, E., NEEDHAM, D. M. and CRAINICEANU, C. M. (2015). Cox regression models with functional covariates for survival data. *Stat. Model.* **15** 256–278. MR3349796 https://doi.org/10.1177/1471082X14565526

GILLIES, R. J., KINAHAN, P. E. and HRICAK, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology* **278** 563–577. https://doi.org/10.1148/radiol.2015151169

GROVE, O., BERGLUND, A. E., SCHABATH, M. B., AERTS, H. J., DEKKER, A., WANG, H., VELAZQUEZ, E. R., LAMBIN, P., GU, Y. et al. (2015). Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS ONE* **10** e0118261.

HAMMOUD, M. A., SAWAYA, R., SHI, W., THALL, P. F. and LEEDS, N. E. (1996). Prognostic significance of preoperative MRI scans in glioblastoma multiforme. *J. Neurooncol.* **27** 65–73. https://doi.org/10.1007/BF00146086

HAO, M., LIU, K., XU, W. and ZHAO, X. (2021). Semiparametric inference for the functional Cox model. *J. Amer. Statist. Assoc.* **116** 1319–1329. MR4309275 https://doi.org/10.1080/01621459.2019.1710155

HARALICK, R. M., SHANMUGAM, K. and DINSTEIN, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3** 610–621.

HAVAEI, M., DAVY, A., WARDE-FARLEY, D., BIARD, A., COURVILLE, A., BENGIO, Y., PAL, C., JODOIN, P.-M. and LAROCHELLE, H. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **35** 18–31. https://doi.org/10.1016/j.media.2016.05.004

HENSON, J. W., GAVIANI, P. and GONZALEZ, R. G. (2005). MRI in treatment of adult gliomas. *Lancet Oncol.* **6** 167–175. https://doi.org/10.1016/S1470-2045(05)01767-5

KARHUNEN, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae Ser. A. I. Math.-Phys.* **1947** 79. MR0023013

KILDAY, J., PALMIERI, F. and FOX, M. D. (1993). Classifying mammographic lesions using computerized image analysis. *IEEE Trans. Med. Imag.* **12** 664–669. https://doi.org/10.1109/42.251116

KONG, D., IBRAHIM, J. G., LEE, E. and ZHU, H. (2018). FLCRM: Functional linear Cox regression model. *Biometrics* **74** 109–117. MR3777931 https://doi.org/10.1111/biom.12748

KUSANO, G., HIRAOKA, Y. and FUKUMIZU, K. (2016). Persistence weighted Gaussian kernel for topological data analysis. *Proceedings of The* 33*rd International Conference on Machine Learning* **48** 2004–2013.

LAWSON, P., SHOLL, A., BROWN, J., FASY, B. T. and WENK, C. (2019). Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Sci. Rep.* **9** 1139.

LEE, E., ZHU, H., KONG, D., WANG, Y., SULLIVAN GIOVANELLO, K. and IBRAHIM, J. G. (2015). BFLCRM: A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease. *Ann. Appl. Stat.* **9** 2153–2178. MR3456370 https://doi.org/10.1214/15-AOAS879

LEVINE, A. B., SCHLOSSER, C., GREWAL, J., COOPE, R., JONES, S. J. and YIP, S. (2019). Rise of the machines: Advances in deep learning for cancer diagnosis. *Trends in Cancer* **5** 157–169.

LI, K., XIAO, J., YANG, J., LI, M., XIONG, X., NIAN, Y., QIAO, L., WANG, H., ERESEN, A. et al. (2019). Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma. *American Journal of Translational Research* **11** 4491.

LOÈVE, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *Revue Sci.* **84** 159–162. MR0017892

MADABHUSHI, A. and LEE, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med. Image Anal.* **33** 170–175.

MATSUDA, T. and MACHII, R. (2015). Morphological distribution of lung cancer from cancer incidence in five continents Vol. X. *Japanese Journal of Clinical Oncology* **45** 404–404.

MILLER, T. R., PINKUS, E., DEHDASHTI, F. and GRIGSBY, P. W. (2003). Improved prognostic value of 18F-FDG PET using a simple visual analysis of tumor characteristics in patients with cervical cancer. *J. Nucl. Med.* **44** 192–197.

MOON, C., LI, Q. and XIAO, G. (2023). Supplement to "Using persistent homology topological features to characterize medical Images: Case studies on lung and brain cancers." https://doi.org/10.1214/22-AOAS1714SUPPA, https://doi.org/10.1214/22-AOAS1714SUPPB

MOON, H.-G., KIM, N., JEONG, S., LEE, M., MOON, H., KIM, J., YOO, T.-K., LEE, H.-B., KIM, J. et al. (2015). The clinical significance and molecular features of the spatial tumor shapes in breast cancers. *PLoS ONE* **10** e0143811.

NELSON, J. S., TSUKADA, Y., SCHOENFELD, D., FULLING, K., LAMARCHE, J. and PERESS, N. (1983). Necrosis as a prognostic criterion in malignant supratentorial, astrocytic gliomas. *Cancer* **52** 550–554. https://doi.org/10.1002/1097-0142(19830801)52:3<550::aid-cncr2820520327>3.0.co;2-c

OBAYASHI, I., HIRAOKA, Y. and KIMURA, M. (2018). Persistence diagrams with linear machine learning models. *J. Appl. Comput. Topol.* **1** 421–449. MR3975560 https://doi.org/10.1007/s41468-018-0013-5

OYAMA, A., HIRAOKA, Y., OBAYASHI, I., SAIKAWA, Y., FURUI, S., SHIRAISHI, K., KUMAGAI, S., HAYASHI, T. and KOTOKU, J. (2019). Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional T1-weighted MR images with a radiomics approach. *Sci. Rep.* **9** 1–10.

QAISER, T., SIRINUKUNWATTANA, K., NAKANE, K., TSANG, Y.-W., EPSTEIN, D. and RAJPOOT, N. (2016). Persistent homology for fast tumor segmentation in whole slide histology images. *Proc. Comput. Sci.* **90** 119–124.

QU, S., WANG, J.-L. and WANG, X. (2016). Optimal estimation for the functional Cox model. *Ann. Statist.* **44** 1708–1738. MR3519938 https://doi.org/10.1214/16-AOS1441

RAZA, S. M., LANG, F. F., AGGARWAL, B. B., FULLER, G. N., WILDRICK, D. M. and SAWAYA, R. (2002). Necrosis and glioblastoma: A friend or a foe? A review and a hypothesis. *Neurosurgery* **51** 2–13. https://doi.org/10.1097/00006123-200207000-00002

REININGHAUS, J., HUBER, S. M., BAUER, U. and KWITT, R. (2015). A stable multi-scale kernel for topological machine learning. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* 4741–4748.

RIZZO, S., BOTTA, F., RAIMONDI, S., ORIGGI, D., FANCIULLO, C., MORGANTI, A. G. and BELLOMI, M. (2018). Radiomics: The facts and the challenges of image analysis. *European Radiology Experimental* **2** 1–8.

ROBINS, V., SAADATFAR, M., DELGADO-FRIEDRICHS, O. and SHEPPARD, A. P. (2016). Percolating length scales from topological persistence analysis of micro-CT images of porous materials. *Water Resour. Res.* **52** 315–329.

SCARPACE, L., MIKKELSEN, L., CHA, T., RAO, S., TEKCHANDANI, S., GUTMAN, S. and PIERCE, D. (2016). Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection. *The Cancer Imaging Archive* **11** 1.

SEFIDGAR, M., SOLTANI, M., RAAHEMIFAR, K., BAZMARA, H., NAYINIAN, S. M. M. and BAZARGAN, M. (2014). Effect of tumor shape, size, and tissue transport properties on drug delivery to solid tumors. *J. Biol. Eng.* **8** 12. https://doi.org/10.1186/1754-1611-8-12

SIEGEL, R. L., MILLER, K. D. and JEMAL, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* **70** 7–30. https://doi.org/10.3322/caac.21590

SOLTANI, M. and CHEN, P. (2012). Effect of tumor shape and size on drug delivery to solid tumors. *Journal of Biological Engineering* **6** 4.

SOMASUNDARAM, E., LITZLER, A., WADHWA, R., OWEN, S. and SCOTT, J. (2021). Persistent homology of tumor CT scans is associated with survival in lung cancer. *Med. Phys.* **48** 7043–7051. https://doi.org/10.1002/mp.15255

SURAWICZ, T. S., MCCARTHY, B. J., KUPELIAN, V., JUKICH, P. J., BRUNER, J. M. and DAVIS, F. G. (1999). Descriptive epidemiology of primary brain and CNS tumors: Results from the Central Brain Tumor Registry of the United States, 1990–1994. *Neuro-Oncol.* **1** 14–25. https://doi.org/10.1093/neuonc/1.1.14

THE CANCER GENOME ATLAS RESEARCH NETWORK (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061.

UPADHYAY, N. and WALDMAN, A. (2011). Conventional MRI evaluation of gliomas. *Br. J. Radiol.* **84** S107–S111.

VOGL, T. J., WORST, T. S., NAGUIB, N. N. N., ACKERMANN, H., GRUBER-ROUH, T. and NOUR-ELDIN, N.-E. A. (2013). Factors influencing local tumor control in patients with neoplastic pulmonary nodules treated with microwave ablation: A risk-factor analysis. *Am. J. Roentgenol.* **200** 665–672. https://doi.org/10.2214/AJR.12.8721

WAGNER, H., CHEN, C. and VUÇINI, E. (2012). Efficient computation of persistent homology for cubical data. In *Topological Methods in Data Analysis and Visualization II. Math. Vis.* 91–106. Springer, Heidelberg. MR3025945 https://doi.org/10.1007/978-3-642-23175-9_7

WANG, S., CHEN, A., YANG, L., CAI, L., XIE, Y., FUJIMOTO, J., GAZDAR, A. and XIAO, G. (2018). Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Sci. Rep.* **8** 10393.

WANG, S., YANG, D. M., RONG, R., ZHAN, X. and XIAO, G. (2019). Pathology image analysis using segmentation deep learning algorithms. *Am. J. Pathol.* **189** 1686–1698.

WANG, B., SUDIJONO, T., KIRVESLAHTI, H., GAO, T., BOYER, D. M., MUKHERJEE, S. and CRAWFORD, L. (2021). A statistical pipeline for identifying physical features that differentiate classes of 3D shapes. *Ann. Appl. Stat.* **15** 638–661. MR4298966 https://doi.org/10.1214/20-aoas1430

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561 https://doi.org/10.1198/016214504000001745

YOKOYAMA, I., SHEAHAN, D. G., CARR, B., KAKIZOE, S., SELBY, R., TZAKIS, A. G., TODO, S., IWATSUKI, S. and STARZL, T. E. (1991). Clinicopathologic factors affecting patient survival and tumor recurrence after orthotopic liver transplantation for hepatocellular carcinoma. *Transplant. Proc.* **23** 2194–2196.

ZAPPA, C. and MOUSA, S. A. (2016). Non-small cell lung cancer: Current treatment and future advances. *Transl. Lung Cancer Res.* **5** 288–300. https://doi.org/10.21037/tlcr.2016.06.07

ZHANG, C., XIAO, G., MOON, C., CHEN, M. and LI, Q. (2020). Bayesian landmark-based shape analysis of tumor pathology images. ArXiv Preprint. Available at arXiv:2012.01149.

ZHU, X., LI, K., KAMALY-ASL, I., CHECKLEY, D., TESSIER, J., WATERTON, J. and JACKSON, A. (2000). Quantification of endothelial permeability, leakage space, and blood volume in brain tumors using combined T1 and T2* contrast-enhanced dynamic MR imaging. *Journal of Magnetic Resonance Imaging*: *An Official Journal of the International Society for Magnetic Resonance in Medicine* **11** 575–585.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2212–2235
https://doi.org/10.1214/22-AOAS1715

BAYESIAN COMBINATORIAL MULTISTUDY FACTOR ANALYSIS

BY ISABELLA N. GRABSKI[1,a], ROBERTA DE VITO[2,b], LORENZO TRIPPA[3,c] AND GIOVANNI PARMIGIANI[3,d]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*isabellagrabski@g.harvard.edu*
[2]*Department of Biostatistics and Data Science Initiative, Brown University,* [b]*roberta_devito@brown.edu*
[3]*Department of Data Science, Dana Farber Cancer Institute,* [c]*ltrippa@jimmy.harvard.edu,* [d]*gp@jimmy.harvard.edu*

Mutations in the *BRCA1* and *BRCA2* genes are known to be highly associated with breast cancer. Identifying both shared and unique transcript expression patterns in blood samples from these groups can shed insight into if and how the disease mechanisms differ among individuals by mutation status, but this is challenging in the high-dimensional setting. A recent method, Bayesian multistudy factor analysis (BMSFA), identifies latent factors common to all studies (or equivalently, groups) and latent factors specific to individual studies. However, BMSFA does not allow for factors shared by more than one but less than all studies. This is critical in our context, as we may expect some but not all signals to be shared by BRCA1- and BRCA2-mutation carriers but not necessarily other high-risk groups. We extend BMSFA by introducing a new method, Tetris, for Bayesian combinatorial multistudy factor analysis which identifies latent factors that any combination of studies or groups can share. We model the subsets of studies that share latent factors with an Indian buffet process and offer a way to summarize uncertainty in the sharing patterns using credible balls. We test our method with an extensive range of simulations and showcase its utility not only in dimension reduction but also in covariance estimation. When applied to transcript expression data from high-risk families grouped by mutation status, Tetris reveals the features and pathways characterizing each group and the sharing patterns among them. Finally, we further extend Tetris to discover groupings of samples when group labels are not provided which can elucidate additional structure in these data.

REFERENCES

ABDI, H. (2007). RV coefficient and congruence coefficient. In *Encyclopedia of Measurement and Statistics* 849–853.

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065 https://doi.org/10.2307/2291387

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

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **38** 65–134. IMS, Beachwood, OH. With discussion by M. Clyde, Dean P. Foster, and Robert A. Stine, and a rejoinder by the authors. MR2000752 https://doi.org/10.1214/lnms/1215540964

DA ROCHA, A. A., GIORGI, R. R., DE SA, S. V., CORREA-GIANNELLA, M. L., FORTES, M. A., CAVALEIRO, A. M., MACHADO, M. C., CESCATO, V. A., BRONSTEIN, M. D. et al. (2006). Hepatocyte growth factor-regulated tyrosine kinase substrate (HGS) and guanylate kinase 1 (GUK1) are differentially expressed in GH-secreting adenomas. *Pituitary* **9** 83–92. https://doi.org/10.1007/s11102-006-9277-1

DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2021). Bayesian multistudy factor analysis for high-throughput biological data. *Ann. Appl. Stat.* **15** 1723–1741. MR4355073 https://doi.org/10.1214/21-aoas1456

Key words and phrases. Multistudy analysis, unsupervised learning, Gibbs sampling, factor analysis, dimension reduction.

DOSHI-VELEZ, F. et al. (2009). The Indian buffet process: Scalable inference and extensions. Master's Thesis, Univ. Cambridge.

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

GHAHRAMANI, Z. and GRIFFITHS, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems* 475–482.

GHAHRAMANI, Z., HINTON, G. E. et al. (1996). The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, Univ. Toronto.

GRABSKI, I. N., DE VITO, R., TRIPPA, L. and PARMIGIANI, G. (2023). Supplement to "Bayesian Combinatorial multistudy factor analysis." https://doi.org/10.1214/22-AOAS1715SUPPA, https://doi.org/10.1214/22-AOAS1715SUPPB

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1998). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models* **335** 77–83. Citeseer.

HORNIK, K. (2005). A CLUE for CLUster ensembles. *J. Stat. Softw.* **14** 1–25.

JOHNSON, S. G. (2020). The NLopt nonlinear-optimization package.

KNOWLES, D. and GHAHRAMANI, Z. (2007). Infinite sparse factor analysis and infinite independent components analysis. In *International Conference on Independent Component Analysis and Signal Separation* 381–388. Springer.

LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. MR1038245 https://doi.org/10.1007/BF01589116

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

LORENZO-SEVA, U. and TEN BERGE, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* **2** 57–64.

MARCHIONNI, L. (2013). The RTopper package: Perform run gene set enrichment across genomic platforms.

MCNICHOLAS, P. D. and MURPHY, T. B. (2008). Parsimonious Gaussian mixture models. *Stat. Comput.* **18** 285–296. MR2413385 https://doi.org/10.1007/s11222-008-9056-0

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

NOCEDAL, J. (1980). Updating quasi-Newton matrices with limited storage. *Math. Comp.* **35** 773–782. MR0572855 https://doi.org/10.2307/2006193

OJI, Y., TATSUMI, N., FUKUDA, M., NAKATSUKA, S.-I., AOYAGI, S., HIRATA, E., NANCHI, I., FUJIKI, F., NAKAJIMA, H. et al. (2014). The translation elongation factor eEF2 is a novel tumor-associated antigen overexpressed in various types of cancers. *Int. J. Oncol.* **44** 1461–1469.

POULIOT, M.-C., KOTHARI, C., JOLY-BEAUPARLANT, C., LABRIE, Y., OUELLETTE, G., SIMARD, J., DROIT, A. and DUROCHER, F. (2017). Transcriptional signature of lymphoblastoid cell lines of BRCA1, BRCA2 and non-BRCA1/2 high risk breast cancer families. *Oncotarget* **8** 78691–78712. https://doi.org/10.18632/oncotarget.20219

R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. and SMYTH, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** e47–e47.

ROY, A., LAVINE, I., HERRING, A. H. and DUNSON, D. B. (2021). Perturbed factor analysis: Accounting for group differences in exposure profiles. *Ann. Appl. Stat.* **15** 1386–1404. MR4316654 https://doi.org/10.1214/20-aoas1435

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

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. With discussion and a reply by the authors. MR3807860 https://doi.org/10.1214/17-BA1073

ZHANG, Y., PARMIGIANI, G. and JOHNSON, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinform.* **2** 078.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2236–2255
https://doi.org/10.1214/22-AOAS1716

USING PROXIES TO IMPROVE FORECAST EVALUATION

BY HAJO HOLZMANN[1,a] AND BERNHARD KLAR[2,b]

[1]*Fachbereich Mathematik und Informatik, Philipps-Universität Marburg,* [a]*holzmann@mathematik.uni-marburg.de*
[2]*Institut für Stochastik, Karlsruher Institut für Technologie (KIT),* [b]*bernhard.klar@kit.edu*

Comparative evaluation of forecasts of statistical functionals relies on comparing averaged losses of competing forecasts after the realization of the quantity Y, on which the functional is based, has been observed. Motivated by high-frequency finance, in this paper we investigate how proxies $\tilde{Y}$ for Y—say volatility proxies—which are observed together with Y can be utilized to improve forecast comparisons. We extend previous results on robustness of loss functions for the mean to general moments and ratios of moments, and show in terms of the variance of differences of losses that using proxies will increase the power in comparative forecast tests. These results apply both to testing conditional as well as unconditional dominance. Finally, we numerically illustrate the theoretical results, both for simulated high-frequency data as well as for high-frequency log returns of several cryptocurrencies.

REFERENCES

AMAYA, D., CHRISTOFFERSEN, P., JACOBS, K. and VASQUEZ, A. (2015). Does realized skewness predict the cross-section of equity returns? *J. Financ. Econ.* **118** 135–167.

BARNDORFF-NIELSEN, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Stat.* **24** 1–13. MR1436619 https://doi.org/10.1111/1467-9469.00045

BROOKS, C., BURKE, S. P., HERAVI, S. and PERSAND, G. (2005). Autoregressive conditional kurtosis. *J. Financ. Econom.* **3** 399–421.

CATANIA, L., GRASSI, S. and RAVAZZOLO, F. (2018). Predicting the volatility of cryptocurrency time-series. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance* (M. Corazza, M. Durbán, A. Grané, C. Perna and M. Sibillo, eds.). Springer, Berlin.

CHU, J., CHAN, S., NADARAJAH, S. and OSTERRIEDER, J. (2017). Garch modelling of cryptocurrencies. *J. Financ. Risk Manag.* **10**.

CORSI, F. (2009). A simple approximate long-memory model of realized volatility. *J. Financ. Econom.* **7** 174–196.

DIEBOLD, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *J. Bus. Econom. Statist.* **33** 1–9. MR3303732 https://doi.org/10.1080/07350015.2014.983236

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **33** 253–263.

DIKS, C., PANCHENKO, V. and VAN DIJK, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econometrics* **163** 215–230. MR2812867 https://doi.org/10.1016/j.jeconom.2011.04.001

DING, Z., GRANGER, C. and ENGLE, R. (1993). A long memory property of stock market returns and a new model. *J. Empir. Finance* **83** 83–106.

EHM, W., GNEITING, T., JORDAN, A. and KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505–562. MR3506792 https://doi.org/10.1111/rssb.12154

FISSLER, T., ZIEGEL, J. F. and GNEITING, T. (2016). Expected shortfall is jointly elicitable with value at risk—implications for backtesting. *Risk Magazine* 58–61.

GHALANOS, A. (2020). rugarch: Univariate GARCH models. R package version 1.4-4.

GIACOMINI, R. and WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578. MR2268409 https://doi.org/10.1111/j.1468-0262.2006.00718.x

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988 https://doi.org/10.1198/jasa.2011.r10138

Key words and phrases. Cryptocurrencies, DM test, forecast comparison, volatility proxies.

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512 https://doi.org/10.1198/jbes.2010.08110
GYAMERAH, S. (2019). Modelling the volatility of bitcoin returns using GARCH models. *Quantitative Finance and Economics* **3** 739–53.
HANSEN, P. R. and LUNDE, A. (2006). Consistent ranking of volatility models. *J. Econometrics* **131** 97–121. MR2275996 https://doi.org/10.1016/j.jeconom.2005.01.005
HARVEY, C. R. and SIDDIQUE, A. (1999). Autoregressive conditional skewness. *J. Financ. Quant. Anal.* **34** 465–487.
HARVEY, C. R. and SIDDIQUE, A. (2000). Conditional skewness in asset pricing tests. *J. Finance* **55** 1263–1295.
HOGA, Y. and DIMITRIADIS, T. (2022). On testing equal conditional predictive ability under measurement error. *J. Bus. Econom. Statist.* **0** 1–13.
HOLZMANN, H. and EULERT, M. (2014). The role of the information set for forecasting—with applications to risk management. *Ann. Appl. Stat.* **8** 595–621. MR3192004 https://doi.org/10.1214/13-AOAS709
HOLZMANN, H. and KLAR, B. (2023). Supplement to "Using proxies to improve forecast evaluation." https://doi.org/10.1214/22-AOAS1716SUPP
KATSIAMPA, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Econom. Lett.* **158** 3–6. MR3681256 https://doi.org/10.1016/j.econlet.2017.06.023
KLEEN, O. (2021). Measurement error sensitivity of loss functions for distribution forecasts. Available at SSRN 3476461.
KOOPMAN, S. J., JUNGBACKER, B. and HOL, E. (2005). Forecasting daily variability of the s&p 100 stock index using historical, realised and implied volatility measurements. *J. Empir. Finance* **12** 445–475.
LAMBERT, P. and LAURENT, S. (2002). Modeling skewness dynamics in series of financial data using skewed location-scale distributions. Working Paper, Université Catholique de Louvain and Université de Liège.
LAU, C. (2015). A simple normal inverse Gaussian-type approach to calculate value-at-risk based on realized moments. *J. Risk* **17** 1–18.
LAURENT, S., ROMBOUTS, J. V. K. and VIOLANTE, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *J. Econometrics* **173** 1–10. MR3019678 https://doi.org/10.1016/j.jeconom.2012.08.004
LEE, G. J. and ENGLE, R. F. (1999). A permanent and transitory component model of stock return volatility. In *Cointegration, Causality and Forecasting: A Festschrift in Honor of Clive W. J. Granger* 980–996.
LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. and GNEITING, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32** 106–127. MR3634309 https://doi.org/10.1214/16-STS588
LI, J. and PATTON, A. J. (2018). Asymptotic inference about predictive accuracy using high frequency data. *J. Econometrics* **203** 223–240. MR3770823 https://doi.org/10.1016/j.jeconom.2017.10.005
NAIMY, V. and HAYEK, M. (2018). Modelling and predicting the bitcoin volatility using garch models. *Int. J. Math. Model. Numer. Optim.* **8** 197–215.
NAIMY, V., HADDAD, O., FERNÁNDEZ-AVILÉS, G. and EL KHOURY, R. (2021). The predictive capacity of GARCH-type models in measuring the volatility of crypto and world currencies. *PLoS ONE* **16** e0245904.
NELSON, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59** 347–370. MR1097532 https://doi.org/10.2307/2938260
NEUBERGER, A. (2012). Realized skewness. *Rev. Financ. Stud.* **25** 3423–3455.
NOLDE, N. and ZIEGEL, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.* **11** 1833–1874. MR3743276 https://doi.org/10.1214/17-AOAS1041
PATTON, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics* **160** 246–256. MR2745881 https://doi.org/10.1016/j.jeconom.2010.03.034
PATTON, A. J. (2020). Comparing possibly misspecified forecasts. *J. Bus. Econom. Statist.* **38** 796–809. MR4154889 https://doi.org/10.1080/07350015.2019.1585256
PATTON, A. J. and SHEPPARD, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series* 801–838. Springer, Berlin.
R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801. MR0331571
SHEN, K., YAO, J. and LI, W. K. (2018). On the surprising explanatory power of higher realized moments in practice. *Stat. Interface* **11** 153–168. MR3690806 https://doi.org/10.4310/SII.2018.v11.n1.a13
STEINWART, I., PASIN, C., WILLIAMSON, R. and ZHANG, S. (2014). Elicitation and identification of properties. In *Conference on Learning Theory* 482–526. PMLR.
WUERTZ, D., SETZ, T., CHALABI, Y., BOUDT, C., CHAUSSE, P. and MIKLOVAC, M. (2020). fGarch: Rmetrics—Autoregressive Conditional Heteroskedastic Modelling. R package version 3042.83.2.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2256–2278
https://doi.org/10.1214/22-AOAS1717

BAYESIAN NONPARAMETRIC MIXTURE MODELING FOR TEMPORAL DYNAMICS OF GENDER STEREOTYPES

BY MARIA DE IORIO[1,a], STEFANO FAVARO[2,b], ALESSANDRA GUGLIELMI[3,c] AND LIFENG YE[4,d]

[1]*Yong Loo Lin School of Medicine, National University of Singapore; ,* [a]*mdi@nus.edu.sg*
[2]*Department of Economics and Statistics, Università di Torino and Collegio Carlo Alberto,* [b]*stefano.favaro@unito.it*
[3]*Department of Mathematics, Politecnico di Milano,* [c]*alessandra.guglielmi@polimi.it*
[4]*Department of Statistical Science, University College London,* [d]*lifeng.ye.13@ucl.ac.uk*

The study of temporal dynamics of gender and ethnic stereotypes is an important topic in many disciplines at the intersection between statistics and social sciences. In this paper we make use of word "embeddings," a common tool in natural language processing and of Bayesian nonparametric mixture modeling for the analysis of temporal dynamics of gender stereotypes in adjectives and occupation over the 20th and 21st centuries in the United States. Our Bayesian nonparametric approach relies on a novel dependent Dirichlet process prior, and it allows for both dynamic density estimation and dynamic clustering of adjective embedding and occupation embedding biases in a hierarchical setting. Posterior inference is performed through a particle Markov chain Monte Carlo algorithm, which is simple and computationally efficient. An application to time-dependent data for adjective embedding bias and for occupation embedding bias shows that our approach enables the quantification of historical trends of gender stereotypes and hence allows to identify how specific adjectives and occupations have become more closely associated with a female rather than male over time.

REFERENCES

ALTMAN, M. (2003). Beyond trashiness: The sexual language of 1970s feminist fiction. *J. Int. Women's Stud.* **4** 7–19.

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

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

ARBEL, J., MENGERSEN, K. and ROUSSEAU, J. (2016). Bayesian nonparametric dependent model for partially replicated data: The influence of fuel spills on species diversity. *Ann. Appl. Stat.* **10** 1496–1516. MR3553233 https://doi.org/10.1214/16-AOAS944

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7** 277–309. MR2934952 https://doi.org/10.1214/12-BA709

BASOW, S. A. (1992). *Gender: Stereotypes and Roles*. Thomson Brooks/Cole Publ., Belmont, CA.

BASSETTI, F., CASARIN, R. and LEISEN, F. (2014). Beta-product dependent Pitman–Yor processes for Bayesian inference. *J. Econometrics* **180** 49–72. MR3188911 https://doi.org/10.1016/j.jeconom.2014.01.007

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38. MR0501592 https://doi.org/10.1093/biomet/65.1.31

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

BOLTANSKI, L. and CHIAPELLO, E. (2006). *The New Spirit of Capitalism*. Verso, London, UK.

BOLUKBASI, T., CHANG, K. W., ZOU, J. Y., SALIGRAMA, V. and KALAI, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* 29, *Barcelona*.

Key words and phrases. Autoregressive models, Bayesian nonparametrics, dependent Dirichlet processes, dynamic density estimation and clustering, gender stereotypes, mixture modeling, particle Markov chain Monte Carlo, word embeddings.

CALISKAN, A., BRYSON, J. J. and NARAYANAN, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* **356** 183–186. https://doi.org/10.1126/science.aal4230
CARON, F., DAVY, M. and DOUCET, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, *Vancouver*.
CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. and VANHEEGHE, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Trans. Signal Process.* **56** 71–84. MR2439814 https://doi.org/10.1109/TSP.2007.900167
COATES, J. (2016). *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. Routledge, London.
COLLOBERT, R., WESTON, J., BOTTOU, L., KARLEN, M., KAVUKCUOGLU, K. and KUKSA, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12** 2493–2537.
DEYOREO, M. and KOTTAS, A. (2018). Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California. *J. Amer. Statist. Assoc.* **113** 68–80. MR3803440 https://doi.org/10.1080/01621459.2017.1328357
DE IORIO, M., FAVARO, S., GUGLIELMI, A. and YE, L. (2023). Supplement to "Bayesian nonparametric mixture modeling for temporal dynamics of gender stereotypes." https://doi.org/10.1214/22-AOAS1717SUPP
DI LUCCA, M. A., GUGLIELMI, A., MÜLLER, P. and QUINTANA, F. A. (2013). A simple class of Bayesian nonparametric autoregression models. *Bayesian Anal.* **8** 63–87. MR3036254 https://doi.org/10.1214/13-BA803
DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586 https://doi.org/10.1093/biomet/asn012
DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 163–183. MR2325270 https://doi.org/10.1111/j.1467-9868.2007.00582.x
FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949
GARG, N., SCHIEBINGER, L., JURAFSKY, D. and ZOU, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* **115** E3635–E3644.
GRIFFIN, J. E. and STEEL, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 179–194. MR2268037 https://doi.org/10.1198/016214505000000727
GRIFFIN, J. E. and STEEL, M. F. J. (2011). Stick-breaking autoregressive processes. *J. Econometrics* **162** 383–396. MR2795625 https://doi.org/10.1016/j.jeconom.2011.03.001
GUOLO, A. and VARIN, C. (2014). Beta regression for time series analysis of bounded data, with application to Canada Google® Flu Trends. *Ann. Appl. Stat.* **8** 74–88. MR3191983 https://doi.org/10.1214/13-AOAS684
GUTIÉRREZ, L., MENA, R. H. and RUGGIERO, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis. *Comput. Statist. Data Anal.* **95** 161–175. MR3425946 https://doi.org/10.1016/j.csda.2015.10.002
HAMILTON, W. L., LESKOVEC, J. and JURAFSKY, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.
HEWITT, N. A. (2012). Feminist frequencies: Regenerating the wave metaphor. *Fem. Stud.* **38** 658–680.
HJORT, N. L. (2000). Bayesian analysis for a generalised Dirichlet process prior. Technical Report, MatematiskInstitutt, Universitetet i Oslo.
HOLMES, J. and MEYERHOFF, M. (2008). *The Handbook of Language and Gender* **25**. Wiley, New York.
HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
LEWIS, M., COOPER BORKENHAGEN, M., CONVERSE, E., LUPYAN, G. and SEIDENBERG, M. S. (2020). What might books be teaching young children about gender? *Psychol. Sci.* 09567976211024643.
LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519 https://doi.org/10.1214/aos/1176346412
MACEACHERN, S. N. (2000). Dependent dirichlet processes. Unpublished Manuscript, Department of Statistics, The Ohio State University, 1–40.
NICHOLSON, L. (2010). Feminism in 'Waves': Useful metaphor or not? *New Polit.* **12** 34–39.
NIETO-BARAJAS, L. E., MÜLLER, P., JI, Y., LU, Y. and MILLS, G. B. (2012). A time-series DDP for functional proteomics profiles. *Biometrics* **68** 859–868. MR3055190 https://doi.org/10.1111/j.1541-0420.2011.01724.x
PAGE, G. L., QUINTANA, F. A. and DAHL, D. B. (2022). Dependent modeling of temporal sequences of random partitions. *J. Comput. Graph. Statist.* **31** 614–627. MR4425090 https://doi.org/10.1080/10618600.2021.1987255
PATI, D., DUNSON, D. B. and TOKDAR, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116** 456–472. MR3049916 https://doi.org/10.1016/j.jmva.2013.01.011
RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. MR2781811 https://doi.org/10.1214/11-BA605
RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2010). Latent stick-breaking processes. *J. Amer. Statist. Assoc.* **105** 647–659. MR2724849 https://doi.org/10.1198/jasa.2010.tm08241

RODRIGUEZ, A. and TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Anal*. **3** 339–365. MR2407430 https://doi.org/10.1214/08-BA313

ROSEN, R. (2013). *The World Split Open*: *How the Modern Women's Movement Changed America*. Tantor eBooks.

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J*. **8** 289–317.

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

SIX, B. and ECKES, T. (1991). A closer look at the complex structure of gender stereotypes. *Sex Roles* **24** 57–71.

TADDY, M. A. (2010). Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime. *J. Amer. Statist*. *Assoc*. **105** 1403–1417. MR2796559 https://doi.org/10.1198/jasa.2010.ap09655

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal*. **13** 559–626. MR3807860 https://doi.org/10.1214/17-BA1073

WILLIAMS, J. E. and BEST, D. L. (1990). *Measuring Sex Stereotypes*: *A Multination Study*. Sage Publ., Thousand Oaks, CA.

XIAO, S., KOTTAS, A. and SANSÓ, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *Ann*. *Appl*. *Stat*. **9** 353–382. MR3341119 https://doi.org/10.1214/14-AOAS796

ZHAO, J., WANG, T., YATSKAR, M., ORDONEZ, V. and CHANG, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, *Copenhagen*.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2279–2304
https://doi.org/10.1214/22-AOAS1718

THE BAYESIAN NESTED LASSO FOR MIXED FREQUENCY REGRESSION MODELS

BY SATYAJIT GHOSH[1,a], KSHITIJ KHARE[2,b] AND GEORGE MICHAILIDIS[2,3,c]

[1]*Division of Biometrics II, CDER, U.S. FDA,* [a]*satyajitghosh90@gmail.com*
[2]*Department of Statistics, University of Florida,* [b]*kdkhare@stat.ufl.edu*
[3]*Informatics Institute, University of Florida,* [c]*gmichail@ufl.edu*

Even though many time series are sampled at different frequencies, their joint evolution is usually modeled and analyzed at a common low frequency. The mixed data sampling (MIDAS) framework was developed to enable joint modeling of mixed frequency temporally evolving data with GDP forecasting as a key motivating application. In this paper we develop a fully Bayesian method to jointly estimate both the appropriate lag as well as the regression coefficients in linear models wherein the response is measured at a lower frequency than the predictors. This is accomplished through a novel prior distribution, coined the Bayesian nested lasso (BNL), that leads to principled selection of the lag of the predictors, reduces the effective number of model parameters through sparsity induced by the lasso component and finally incorporates desirable decay patterns over time lags in the magnitude of the corresponding regression coefficients. Further, it is easy to obtain samples from the posterior distribution due to the closed form expressions for the conditional distributions of the model parameters. Numerical results, obtained from synthetic and macroeconomic data, illustrate the good performance of the proposed Bayesian framework in parameter selection and estimation and in the key task of GDP forecasting.

REFERENCES

ANDREOU, E., GHYSELS, E. and KOURTELLOS, A. (2010). Regression models with mixed sampling frequencies. *J. Econometrics* **158** 246–261. MR2720834 https://doi.org/10.1016/j.jeconom.2010.01.004

ARMESTO, M. T., ENGEMANN, K. M., OWYANG, M. T. et al. (2010). Forecasting with mixed frequencies. *Fed. Reserve Bank St. Louis* **92** 521–36.

BABII, A., GHYSELS, E. and STRIAUKAS, J. (2022). Machine learning time series regressions with an application to nowcasting. *J. Bus. Econom. Statist.* **40** 1094–1106. MR4439275 https://doi.org/10.1080/07350015.2021.1899933

BAI, J., GHYSELS, E. and WRIGHT, J. H. (2013). State space models and MIDAS regressions. *Econometric Rev.* **32** 779–813. MR3041103 https://doi.org/10.1080/07474938.2012.690675

BAŃBURA, M., GIANNONE, D. and REICHLIN, L. (2010). Large Bayesian vector auto regressions. *J. Appl. Econometrics* **25** 71–92. MR2751790 https://doi.org/10.1002/jae.1137

BAŃBURA, M., GIANNONE, D., MODUGNO, M. and REICHLIN, L. (2013). Now-casting and the real-time data flow. *Handb. Econom. Forecast.* **2** 195–237.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BYBEE, L., KELLY, B. T., MANELA, A. and XIU, D. (2020). The structure of economic news. Working Paper 26648.

CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2015). Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility. *J. Roy. Statist. Soc. Ser. A* **178** 837–862. MR3405481 https://doi.org/10.1111/rssa.12092

DIEBOLD, F. and MARIANO, R. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

ERAKER, B., CHIU, C. W., FOERSTER, A. T., KIM, T. B. and SEOANE, H. D. (2014). Bayesian mixed frequency VARs. *J. Financ. Econom.* **13** 698–721.

Key words and phrases. Mixed frequency regression, nested lasso, lag selection, spike-and-slab priors.

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
FORONI, C., GUÉRIN, P. and MARCELLINO, M. (2015). Markov-switching mixed-frequency VAR models. *Int. J. Forecast.* **31**.
FORONI, C. and MARCELLINO, M. (2013a). A survey of econometric methods for mixed-frequency data. Working Paper 2013/06.
FORONI, C. and MARCELLINO, M. G. (2013b). A survey of econometric methods for mixed-frequency data. SSRN 2268912.
FORONI, C., MARCELLINO, M. and SCHUMACHER, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *J. Roy. Statist. Soc. Ser. A* **178** 57–82. MR3291761 https://doi.org/10.1111/rssa.12043
GALVÃO, A. B. (2013). Changes in predictive ability with mixed frequency data. *Int. J. Forecast.* **29** 395–410.
GALVÃO, A. B. and OWYANG, M. (2022). Forecasting low-frequency macroeconomic events with high-frequency data. *J. Appl. Econometrics* **37** 1314–1333. MR4523212
GEFANG, D., KOOP, G. and POON, A. (2020). Computationally efficient inference in large Bayesian mixed frequency VARs. *Econom. Lett.* **191** 109120. MR4081477 https://doi.org/10.1016/j.econlet.2020.109120
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
GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.
GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2021). Strong selection consistency of Bayesian vector autoregressive models based on a pseudo-likelihood approach. *Ann. Statist.* **49** 1267–1299. MR4298864 https://doi.org/10.1214/20-aos1992
GHOSH, S., KHARE, K. and MICHAILIDIS, G. (2023). Supplement to "The Bayesian nested lasso for mixed frequency regression models." https://doi.org/10.1214/22-AOAS1718SUPP
GHYSELS, E. and QIAN, H. (2019). Estimating MIDAS regressions via OLS with polynomial parameter profiling. *Econom. Stat.* **9** 1–16. MR3907670 https://doi.org/10.1016/j.ecosta.2018.02.001
GHYSELS, E., SANTA-CLARA, P. and VALKANOV, R. (2004). The MIDAS touch: Mixed data sampling regression models. CIRANO Working Papers.
GHYSELS, E., SINKO, A. and VALKANOV, R. (2007). Midas regressions: Further results and new directions. *Econometric Rev.* **26** 53–90. MR2339264 https://doi.org/10.1080/07474930600972467
GIANNONE, D., REICHLIN, L. and SMALL, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *J. Monet. Econ.* **55** 665–676.
GUÉRIN, P. and MARCELLINO, M. (2013). Markov-switching MIDAS models. *J. Bus. Econom. Statist.* **31** 45–56. MR3030797 https://doi.org/10.1080/07350015.2012.727721
HARVEY, D., LEYBOURNE, S. and NEWBOLD, P. (1997). Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **13** 281–291.
HAVRANEK, T. and RUSNAK, M. (2012). Transmission lags of monetary policy: A meta-analysis. William Davidson Institute Working Paper.
HIGGINS, P. C. (2014). GDPNow: A Model for GDP 'Nowcasting'. FRB Atlanta Working Paper 2014-7.
LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* **2** 245–263. MR2415602 https://doi.org/10.1214/07-AOAS139
MARSILLI, C. (2014). Variable selection in predictive midas models. Working Paper.
MCCRACKEN, M. W., OWYANG, M. and SEKHPOSYAN, T. (2015). Real-time forecasting with a large, mixed frequency, Bayesian VAR. FRB St. Louis Working Paper (2015-30).
MOGLIANI, M. and SIMONI, A. (2021). Bayesian MIDAS penalized regressions: Estimation, selection, and prediction. *J. Econometrics* **222** 833–860. MR4236449 https://doi.org/10.1016/j.jeconom.2020.07.022
NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987 https://doi.org/10.1214/14-AOS1207
RODRIGUEZ, A. and PUGGIONI, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction. *Int. J. Forecast.* **26**.
STOCK, J. H. and WATSON, M. W. (2005). An empirical comparison of methods for forecasting using many predictors. Princeton University. Manuscript.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
UEMATSU, Y. and TANAKA, S. (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *Econom. J.* **22** 34–56. MR4021110 https://doi.org/10.1111/ectj.12117
WEST, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica* **64** 1067–1084. MR1403232 https://doi.org/10.2307/2171956
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

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2305–2325
https://doi.org/10.1214/22-AOAS1719

SPATIAL QUANTILE AUTOREGRESSION FOR SEASON WITHIN YEAR DAILY MAXIMUM TEMPERATURE DATA

BY JORGE CASTILLO-MATEO[1,a], JESÚS ASÍN[1,b], ANA C. CEBRIÁN[1,c], ALAN E. GELFAND[2,e] AND JESÚS ABAURREA[1,d]

[1]*Department of Statistical Methods, University of Zaragoza,* [a]*jorgecm@unizar.es,* [b]*jasin@unizar.es,* [c]*acebrian@unizar.es,* [d]*abaurrea@unizar.es*

[2]*Department of Statistical Science, Duke University,* [e]*alan@stat.duke.edu*

Regression is the most widely used modeling tool in statistics. Quantile regression offers a strategy for enhancing the regression picture beyond customary mean regression. With time-series data, we move to quantile autoregression and, finally, with spatially referenced time series, we move to space-time quantile regression. Here, we are concerned with the spatiotemporal evolution of daily maximum temperature, particularly with regard to extreme heat. Our motivating data set is 60 years of daily summer maximum temperature data over Aragón in Spain. Hence, we work with time on two scales—days within summer season across years—collected at geocoded station locations. For a specified quantile, we fit a very flexible, mixed-effects autoregressive model, introducing four spatial processes. We work with asymmetric Laplace errors to take advantage of the available conditional Gaussian representation for these distributions. Further, while the autoregressive model yields conditional quantiles, we demonstrate how to extract marginal quantiles with the asymmetric Laplace specification. Thus, we are able to interpolate quantiles for any days within years across our study region.

REFERENCES

AEMET (2011). Atlas Climático Ibérico—Iberian Climate Atlas. Ministerio de Medio Ambiente, y Medio Rural y Marino; Agencia Estatal de Meteorología; and Instituto de Meteorologia de Portugal. https://doi.org/10.31978/784-11-002-5

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184 https://doi.org/10.1201/b17115

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CASTILLO-MATEO, J., LAFUENTE, M., ASÍN, J., CEBRIÁN, A. C., GELFAND, A. E. and ABAURREA, J. (2022). Spatial modeling of day-within-year temperature time series: An examination of daily maximum temperatures in Aragón, Spain. *J. Agric. Biol. Environ. Stat.* **27** 487–505. MR4459077 https://doi.org/10.1007/s13253-022-00493-3

CASTILLO-MATEO, J., ASÍN, J., CEBRIÁN, A. C., GELFAND, A. E. and ABAURREA, J. (2023). Supplement to "Spatial quantile autoregression for season within year daily maximum temperature data." https://doi.org/10.1214/22-AOAS1719SUPP

CATTIAUX, J. and RIBES, A. (2018). Defining single extreme weather events in a climate perspective. *Bull. Am. Meteorol. Soc.* **99** 1557–1568. https://doi.org/10.1175/BAMS-D-17-0281.1

CHEN, X. and TOKDAR, S. T. (2021). Joint quantile regression for spatial data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 826–852. MR4320003 https://doi.org/10.1111/rssb.12467

CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. MR1984779 https://doi.org/10.1016/S0304-4076(03)00100-3

DAS, P. and GHOSAL, S. (2017a). Bayesian quantile regression using random B-spline series prior. *Comput. Statist. Data Anal.* **109** 121–143. MR3603645 https://doi.org/10.1016/j.csda.2016.11.014

Key words and phrases. Asymmetric Laplace distribution, Gaussian process, hierarchical model, marginal quantile, Markov chain Monte Carlo, seasonal time series.

DAS, P. and GHOSAL, S. (2017b). Analyzing ozone concentration by Bayesian spatio-temoral quantile regression. *Environmetrics* **28** e2443, 15 pp. MR3660099 https://doi.org/10.1002/env.2443
GAO, M. and FRANZKE, C. L. E. (2017). Quantile regression-based spatiotemporal analysis of extreme temperature change in China. *J. Climate* **30** 9897–9914. https://doi.org/10.1175/JCLI-D-17-0356.1
GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275 https://doi.org/10.1093/biomet/82.3.479
HALLIN, M., LU, Z. and YU, K. (2009). Local linear spatial quantile regression. *Bernoulli* **15** 659–686. MR2555194 https://doi.org/10.3150/08-BEJ168
HAUGEN, M. A., STEIN, M. L., MOYER, E. J. and SRIVER, R. L. (2018). Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. *J. Climate* **31** 8573–8588. https://doi.org/10.1175/JCLI-D-17-0782.1
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. and MACHADO, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.* **94** 1296–1310. MR1731491 https://doi.org/10.2307/2669943
KOENKER, R. and XIAO, Z. (2006). Quantile autoregression. *J. Amer. Statist. Assoc.* **101** 980–990. MR2324109 https://doi.org/10.1198/016214506000000672
KOTZ, S., KOZUBOWSKI, T. J. and PODGÓRSKI, K. (2001). *The Laplace Distribution and Generalizations*: *A Revisit with Applications to Communications*, *Economics*, *Engineering*, *and Finance*. Birkhäuser, Inc., Boston, MA. MR1935481 https://doi.org/10.1007/978-1-4612-0173-1
KOZUMI, H. and KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81** 1565–1578. MR2851270 https://doi.org/10.1080/00949655.2010.496117
LEE, C. C. (2021). Weather whiplash: Trends in rapid temperature changes in a warming climate. *Int. J. Climatol.* **42** 4214–4222. https://doi.org/10.1002/joc.7458
LI, G., LI, Y. and TSAI, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. *J. Amer. Statist. Assoc.* **110** 246–261. MR3338500 https://doi.org/10.1080/01621459.2014.892007
LUM, K. and GELFAND, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Anal.* **7** 235–258. MR2934947 https://doi.org/10.1214/12-BA708
MCKINNON, K. A. and POPPICK, A. (2020). Estimating changes in the observed relationship between humidity and temperature using noncrossing quantile smoothing splines. *J. Agric. Biol. Environ. Stat.* **25** 292–314. MR4132962 https://doi.org/10.1007/s13253-020-00393-4
NAVARRO-SERRANO, F., LÓPEZ-MORENO, J. I., AZORIN-MOLINA, C., ALONSO-GONZÁLEZ, E., TOMÁS-BURGUERA, M., SANMIGUEL-VALLELADO, A., REVUELTO, J. and VICENTE-SERRANO, S. M. (2018). Estimation of near-surface air temperature lapse rates over continental Spain and its mountain areas. *Int. J. Climatol.* **38** 3233–3249. https://doi.org/10.1002/joc.5497
NEELON, B., LI, F., BURGETTE, L. F. and BENJAMIN NEELON, S. E. (2015). A spatiotemporal quantile regression model for emergency department expenditures. *Stat. Med.* **34** 2559–2575. MR3368401 https://doi.org/10.1002/sim.6480
PEÑA-ANGULO, D., GONZALEZ-HIDALGO, J. C., SANDONÍS, L., BEGUERÍA, S., TOMAS-BURGUERA, M., LÓPEZ-BUSTINS, J. A., LEMUS-CANOVAS, M. and MARTIN-VIDE, J. (2021). Seasonal temperature trends on the Spanish mainland: A secular study (1916–2015). *Int. J. Climatol.* **41** 3071–3084. https://doi.org/10.1002/joc.7006
PETERS, G. W. (2018). General quantile time series regressions for applications in population demographics. *Risks* **6** 97. https://doi.org/10.3390/risks6030097
REICH, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 535–553. MR2960737 https://doi.org/10.1111/j.1467-9876.2011.01025.x
REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698 https://doi.org/10.1198/jasa.2010.ap09237
SRIRAM, K., RAMAMOORTHI, R. V. and GHOSH, P. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Anal.* **8** 479–504. MR3066950 https://doi.org/10.1214/13-BA817
TAN, X., GAN, T. Y. and CHEN SHU LIU, B. (2019). Modeling distributional changes in winter precipitation of Canada using Bayesian spatiotemporal quantile regression subjected to different teleconnections. *Clim. Dyn.* **52** 2105–2124. https://doi.org/10.1007/s00382-018-4241-0
TOKDAR, S. T. and KADANE, J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Anal.* **7** 51–72. MR2896712 https://doi.org/10.1214/12-BA702

YANG, Y. and HE, X. (2015). Quantile regression for spatially correlated data: An empirical likelihood approach. *Statist. Sinica* **25** 261–274. MR3328814

YANG, C., LI, L. and XU, J. (2018). Changing temperature extremes based on CMIP5 output via semi-parametric quantile regression approach. *Int. J. Climatol.* **38** 3736–3748. https://doi.org/10.1002/joc.5524

YANG, Y. and TOKDAR, S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *J. Amer. Statist. Assoc.* **112** 1107–1120. MR3735363 https://doi.org/10.1080/01621459.2016.1192545

YANG, Y., WANG, H. J. and HE, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *Int. Stat. Rev.* **84** 327–344. MR3580414 https://doi.org/10.1111/insr.12114

YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. MR1861390 https://doi.org/10.1016/S0167-7152(01)00124-9

ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2326–2344
https://doi.org/10.1214/22-AOAS1720

A DYNAMIC SCREENING ALGORITHM FOR HIERARCHICAL BINARY MARKETING DATA

BY YIMEI FAN[1,a], YUAN LIAO[2,b], ILYA O. RYZHOV[3,c] AND KUNPENG ZHANG[3,d]

[1]*Department of Mathematics, University of Maryland,* [a]*ymfan@math.umd.edu*
[2]*Department of Economics, Rutgers University,* [b]*yuan.liao@rutgers.edu*
[3]*Robert H. Smith School of Business, University of Maryland,* [c]*iryzhov@umd.edu,* [d]*kpzhang@umd.edu*

In many applications of business and marketing analytics, predictive models are fit using hierarchically structured data: common characteristics of products, customers, or web pages are represented as categorical variables, and each category can be split up into multiple subcategories at a lower level of the hierarchy. The model may thus contain hundreds of thousands of binary variables, necessitating the use of variable selection to screen out large numbers of irrelevant or insignificant features. We propose a new dynamic screening method, based on the distance correlation criterion, designed for hierarchical binary data. Our method can screen out large parts of the hierarchy at the higher levels, avoiding the need to explore many lower-level features and greatly reducing the computational cost of screening. The practical potential of the method is demonstrated in a case application on user-brand interaction data from Facebook.

REFERENCES

BACH, F., JENATTON, R., MAIRAL, J. and OBOZINSKI, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.* **27** 450–468. MR3025128 https://doi.org/10.1214/12-STS394

BARUT, E., FAN, J. and VERHASSELT, A. (2016). Conditional sure independence screening. *J. Amer. Statist. Assoc.* **111** 1266–1277. MR3561948 https://doi.org/10.1080/01621459.2015.1092974

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163 https://doi.org/10.3150/11-BEJ410

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131 https://doi.org/10.3982/ECTA9626

BERTSIMAS, D., O'HAIR, A., RELYEA, S. and SILBERHOLZ, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Manage. Sci.* **62** 1511–1531.

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

CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-Normalized Processes*: *Limit Theory and Statistical Applications*. *Probability and Its Applications* (*New York*). Springer, Berlin. MR2488094 https://doi.org/10.1007/978-3-540-85636-8

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009 https://doi.org/10.1214/07-AOS504

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969 https://doi.org/10.1198/jasa.2011.tm09779

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. MR3689312 https://doi.org/10.1111/rssb.12204

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

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

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861 https://doi.org/10.1214/10-AOS798

Key words and phrases. Sure screening, hierarchical data, distance correlation.

FAN, Y., LIAO, Y., RYZHOV, I. O. and ZHANG, K. (2023). Supplement to "A dynamic screening algorithm for hierarchical binary marketing data." https://doi.org/10.1214/22-AOAS1720SUPPA, https://doi.org/10.1214/22-AOAS1720SUPPB

FERREIRA, J. A. and ZWINDERMAN, A. H. (2006). On the Benjamini–Hochberg method. *Ann. Statist.* **34** 1827–1849. MR2283719 https://doi.org/10.1214/009053606000000425

HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697 https://doi.org/10.1080/01621459.2014.881741

HAO, N. and ZHANG, H. H. (2017). A note on high-dimensional linear regression with interactions. *Amer. Statist.* **71** 291–297. MR3750934 https://doi.org/10.1080/00031305.2016.1264311

HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics* **58** 435–447. MR3556612 https://doi.org/10.1080/00401706.2015.1054435

KIM, S. and XING, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the* 27*th International Conference on International Conference on Machine Learning* 543–550.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat.* **6** 1095–1117. MR3012522 https://doi.org/10.1214/12-AOAS549

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677 https://doi.org/10.1111/rssb.12050

LI, J., NETESSINE, S. and KOULAYEV, S. (2018). Price to compete... With many: How to identify price competition in high dimensional space. *Manage. Sci.* **64** 3971–4470.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900 https://doi.org/10.1080/01621459.2012.695654

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046 https://doi.org/10.1214/12-AOS1024

LIU, W. and SHAO, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025. MR3262475 https://doi.org/10.1214/14-AOS1249

QU, H., RYZHOV, I. O., FU, M. C., BERGERSON, E., KURKA, M. and KOPACEK, L. (2020). Learning demand curves in B2B pricing: A new framework and case study. *Production and Operations Management* **29** 1287–1306.

RUDIN, C., WALTZ, D., ANDERSON, R. N., BOULANGER, A., SALLEB-AOUISSI, A., CHOW, M., DUTTA, H., GROSS, P. N., HUANG, B. et al. (2012). Machine learning for the New York City power grid. *IEEE Trans. Pattern Anal. Mach. Intell.* **34** 328–345.

RYZHOV, I. O., HAN, B. and BRADIĆ, J. (2016). Cultivating disaster donors using data analytics. *Manage. Sci.* **62** 849–866.

SMITHSON, M. and MERKLE, E. C. (2013). *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. CRC Press, Boca Raton.

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1233–1303.

SZÉKELY, G. J. and RIZZO, M. L. (2012). On the uniqueness of distance covariance. *Statist. Probab. Lett.* **82** 2278–2282. MR2979766 https://doi.org/10.1016/j.spl.2012.08.007

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

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

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809 https://doi.org/10.1214/009053607000000929

XUE, Z., WANG, Z. and ETTL, M. (2016). Pricing personalized bundles: A new approach and an empirical study. *Manuf. Serv. Oper. Manag.* **18** 51–68.

YAN, X. and BIEN, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statist. Sci.* **32** 531–560. MR3730521 https://doi.org/10.1214/17-STS622

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Amer. Statist. Assoc.* **103** 309–316. MR2420235 https://doi.org/10.1198/016214507000001373

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZHANG, K., BHATTACHARYYA, S. and RAM, S. (2016). Large-scale network analysis for online social brand advertising. *MIS Q.* **40** 849–868.

ZHAO, S. D. and LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105** 397–411. MR2877525 https://doi.org/10.1016/j.jmva.2011.08.002

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566 https://doi.org/10.1214/07-AOS584

ZHOU, J., FOSTER, D. P., STINE, R. A. and UNGAR, L. H. (2006). Streamwise feature selection. *J. Mach. Learn. Res.* **7** 1861–1885. MR2274426

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 https://doi.org/10.1198/jasa.2011.tm10563

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2345–2363
https://doi.org/10.1214/22-AOAS1721

PENALIZED ESTIMATING EQUATIONS FOR GENERALIZED LINEAR MODELS WITH MULTIPLE IMPUTATION

BY YANG LI[1,a], HAOYU YANG[1,b], HAOCHEN YU[1,c], HANWEN HUANG[2,d] AND YE SHEN[2,e]

[1]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [a]*yang.li@ruc.edu.cn,* [b]*haoyuyang@ruc.edu.cn,* [c]*haocheny@ruc.edu.cn*

[2]*Department of Epidemiology and Biostatistics, University of Georgia,* [d]*huanghw@uga.edu,* [e]*yeshen@uga.edu*

Missing values among variables present a challenge in variable selection in the generalized linear model. Common strategies that delete observations with missing information may cause serious information loss. Multiple imputation has been widely used in recent years because it provides unbiased statistical results given a correctly specified imputation model and considers the uncertainty of the missing data. However, variable selection methods in the generalized linear model with multiply-imputed data have not yet been studied widely. In this study, we introduce penalized estimating equations for generalized linear models with multiple imputation (PEE–MI), which incorporates the correlation of multiple imputed observations into the objective function. The theoretical performance of the proposed PEE–MI depends on the penalized function adopted. We use the adaptive least absolute shrinkage and selection operator (adaptive LASSO) as an illustrating example. Simulations show that PEE–MI outperforms the alternatives. The proposed method is shown to select variables with clinical relevance when applied to a database of laboratory-diagnosed A/H7N9 patients in the Zhejiang province, China.

REFERENCES

AZUR, M. J., STUART, E. A., FRANGAKIS, C. and LEAF, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* **20** 40–49. https://doi.org/10.1002/mpr.329

CHARKHI, A. and CLAESKENS, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* **105** 645–664. MR3842890 https://doi.org/10.1093/biomet/asy018

CHEN, Q. and WANG, S. (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Stat. Med.* **32** 3646–3659. MR3095503 https://doi.org/10.1002/sim.5783

CHENG, Q. L., DING, H., SUN, Z., KAO, Q. J., YANG, X. H., HUANG, R. J., WEN, Y. Y., WANG, J. and XIE, L. (2015). Retrospective study of risk factors for mortality in human avian influenza A H7N9 cases in Zhejiang Province, China, March 2013 to June 2014. *Int. J. Infect. Dis.* **39** 95–101.

CHENG, Q., ZHAO, G., XIE, L. and WANG, X. (2018). Impacts of age and gender at the risk of underlying medical conditions and death in patients with avian influenza A H7N9: A meta-analysis study. *Ther. Clin. Risk Manag.* **14** 1615–1626.

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

FANG, E. X., NING, Y. and LI, R. (2020). Test of significance for high-dimensional longitudinal data. *Ann. Statist.* **48** 2622–2645. MR4152115 https://doi.org/10.1214/19-AOS1900

FERRARI, D. and YANG, Y. (2015). Confidence sets for model selection by F-testing. *Statist. Sinica* **25** 1637–1658. MR3409085

GERONIMI, J. and SAPORTA, G. (2017). Variable selection for multiply-imputed data with penalized generalized estimating equations. *Comput. Statist. Data Anal.* **110** 103–114. MR3612611 https://doi.org/10.1016/j.csda.2017.01.001

GOH, G. and KIM, J. K. (2021). Accounting for model uncertainty in multiple imputation under complex sampling. *Scand. J. Stat.* **48** 930–949. MR4303563 https://doi.org/10.1111/sjos.12473

HUANG, J. and MA, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* **16** 176–195. MR2608284 https://doi.org/10.1007/s10985-009-9144-2

Key words and phrases. Missing data, multiple imputation, penalized estimating equations, variable selection.

JIANG, Y., HE, Y. and ZHANG, H. (2016). Variable selection with prior information for generalized linear models via the prior LASSO method. *J. Amer. Statist. Assoc.* **111** 355–376. MR3494665 https://doi.org/10.1080/01621459.2015.1008363

KIM, J. K. and YANG, S. (2017). A note on multiple imputation under complex sampling. *Biometrika* **104** 221–228. MR3626470 https://doi.org/10.1093/biomet/asw058

LEE, D. and KIM, J. K. (2022). Semiparametric imputation using conditional Gaussian mixture models under item nonresponse. *Biometrics* **78** 227–237. MR4408583 https://doi.org/10.1111/biom.13410

LEI, J. (2020). Cross-validation with confidence. *J. Amer. Statist. Assoc.* **115** 1978–1997. MR4189771 https://doi.org/10.1080/01621459.2019.1672556

LI, Y., LUO, Y., FERRARI, D., HU, X. and QIN, Y. (2019). Model confidence bounds for variable selection. *Biometrics* **75** 392–403. MR3999165 https://doi.org/10.1111/biom.13024

LI, Y., YANG, H., YU, H., HUANG, H. and SHEN, Y. (2023). Supplement to "Penalized estimating equations for generalized linear models with multiple imputation." https://doi.org/10.1214/22-AOAS1721SUPP

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LITTLE, R. J. A. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* **793**. John Wiley & Sons. MR1925014 https://doi.org/10.1002/9781119013563

LIU, S., SUN, J., CAI, J., MIAO, Z., LU, M., QIN, S., WANG, X., LV, H., YU, Z. et al. (2013). Epidemiological, clinical and viral characteristics of fatal cases of human avian influenza A H7N9 virus in Zhejiang Province, China. *J. Infect.* **67** 595–605.

LIU, Y., WANG, Y., FENG, Y. and WALL, M. M. (2016). Variable selection and prediction with incomplete high-dimensional data. *Ann. Appl. Stat.* **10** 418–450. MR3480502 https://doi.org/10.1214/15-AOAS899

LONG, Q. and JOHNSON, B. A. (2015). Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics* **16** 596–610. MR3365449 https://doi.org/10.1093/biostatistics/kxv003

LV, J., YANG, H. and GUO, C. (2015). An efficient and robust variable selection method for longitudinal generalized linear models. *Comput. Statist. Data Anal.* **82** 74–88. MR3282167 https://doi.org/10.1016/j.csda.2014.08.006

MARTINEZ, L., CHENG, W., WANG, X., LING, F., MU, L., LI, C., HUO, X., EBELL, M. H., HUANG, H. et al. (2019). A risk classification model to predict mortality among laboratory-confirmed avian influenza A H7N9 patients: A population-based observational cohort study. *J. Infect. Dis.* **220** 1780–1789.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836. MR1813977 https://doi.org/10.1093/biomet/87.4.823

SIDDIQUE, J., HAREL, O., CRESPI, C. M. and HEDEKER, D. (2014). Binary variable multiple-model multiple imputation to address missing data mechanism uncertainty: Application to a smoking cessation trial. *Stat. Med.* **33** 3013–3028. MR3260519 https://doi.org/10.1002/sim.6137

UYEKI, T. M. and COX, N. J. (2013). Global concerns regarding novel influenza A (H7N9) virus infections. *N. Engl. J. Med.* **368** 1862–1864. https://doi.org/10.1056/NEJMp1304661

WANG, L. and MA, W. (2021). Improved empirical likelihood inference and variable selection for generalized linear models with longitudinal nonignorable dropouts. *Ann. Inst. Statist. Math.* **73** 623–647. MR4247073 https://doi.org/10.1007/s10463-020-00761-4

WANG, X., JIANG, H., WU, P., UYEKI, T. M., FENG, L., LAI, S., WANG, L., HUO, X., XU, K. et al. (2017). Epidemiology of avian influenza A H7N9 virus in human beings across five epidemics in mainland China, 2013–17: An epidemiological study of laboratory-confirmed case series. *Lancet Infect. Dis.* **17** 822–832.

XUE, F. and QU, A. (2021). Integrating multisource block-wise missing data in model selection. *J. Amer. Statist. Assoc.* **116** 1914–1927. MR4353722 https://doi.org/10.1080/01621459.2020.1751176

YANG, X., BELIN, T. R. and BOSCARDIN, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61** 498–506. MR2140922 https://doi.org/10.1111/j.1541-0420.2005.00317.x

YANG, Y., LI, X., BIRKHEAD, G. S., ZHENG, Z. and LU, J. H. (2019). Clinical indices and mortality of hospitalized avian influenza A H7N9 patients in Guangdong, China. *Chin. Med. J.* **132** 302–310.

ZHAO, J., YANG, Y. and NING, Y. (2018). Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Statist. Sinica* **28** 2125–2148. MR3839007

ZHENG, C., FERRARI, D. and YANG, Y. (2019). Model selection confidence sets by likelihood ratio testing. *Statist. Sinica* **29** 827–851. MR3931390

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2364–2386
https://doi.org/10.1214/22-AOAS1723

SUBBOTIN GRAPHICAL MODELS FOR EXTREME VALUE DEPENDENCIES WITH APPLICATIONS TO FUNCTIONAL NEURONAL CONNECTIVITY

BY ANDERSEN CHANG[1,a] AND GENEVERA I. ALLEN[2,b]

[1]*Department of Statistics, Rice University,* [a]*atc7@rice.edu*
[2]*Department of Electrical and Computer Engineering, Rice University,* [b]*gallen@rice.edu*

With modern calcium imaging technology, activities of thousands of neurons can be recorded in vivo. These experiments can potentially provide new insights into intrinsic functional neuronal connectivity, defined as contemporaneous correlations between neuronal activities. As a common tool for estimating conditional dependencies in high-dimensional settings, graphical models are a natural choice for estimating functional connectivity networks. However, raw neuronal activity data presents a unique challenge: the relevant information in the data lies in rare extreme value observations that indicate neuronal firing rather than in the observations near the mean. Existing graphical modeling techniques for extreme values rely on binning or thresholding observations which may not be appropriate for calcium imaging data. In this paper we develop a novel class of graphical models, called the Subbotin graphical model, which finds sparse conditional dependency structures with respect to the extreme value observations without requiring data preprocessing. We first derive the form of the Subbotin graphical model and show the conditions under which it is normalizable. We then study the empirical performance of the Subbotin graphical model and compare it to existing extreme value graphical modeling techniques and functional connectivity models from neuroscience through several simulation studies as well as a real-world calcium imaging data example.

REFERENCES

ALI, A., KOLTER, J. Z. and TIBSHIRANI, R. J. (2016). The multiple quantile graphical model. arXiv preprint, arXiv:1607.00515.

ALLEN, G. I. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.

BALI, T. G. (2003). The generalized extreme value distribution. *Econom. Lett.* **79** 423–427.

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

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

BISWAL, B., ZERRIN YETKIN, F., HAUGHTON, V. M. and HYDE, J. S. (1996). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34** 537–541.

BUCKNER, R. L., KRIENEN, F. M. and YEO, B. T. (2013). Opportunities and limitations of intrinsic functional connectivity MRI. *Nat. Neurosci.* **16** 832–837.

BUISHAND, T. A. (1989). Statistics of extremes in climatology. *Stat. Neerl.* **1** 1172–1188.

CAPOBIANCO, R. (2000). Robustness aspects of the generalized normal distribution. *Quad. Stat.* **2** 127–145.

CHANG, A. and ALLEN, G. I. (2023). Supplement to "Subbotin graphical models for extreme value dependencies with applications to functional neuronal connectivity." https://doi.org/10.1214/22-AOAS1723SUPPA, https://doi.org/10.1214/22-AOAS1723SUPPB

CHANG, A., WANG, M. and ALLEN, G. I. (2021). Sparse regression for extreme values. *Electron. J. Stat.* **15** 5995–6035. MR4355702 https://doi.org/10.1214/21-ejs1937

CHANG, A., YAO, T. and ALLEN, G. I. (2019). Graphical models and dynamic latent factors for modeling functional brain connectivity. In 2019 *IEEE Data Science Workshop* (*DSW*) 57–63.

Key words and phrases. Calcium imaging, exponential family graphical models, extreme values, generalized normal distribution, graphical models, Subbotin distribution.

DAHLHAUS, R., EICHLER, M. and SANDKUHLER, J. (1997). Identification of synaptic connections in neural ensembles by graphical models. *J. Neurosci. Methods* **77** 93–107.

DE SIMIONI, S. (1968). Su una estensione delle curve normali di ordina "r" alle variabili doppie. *Statistica* **28** 151–178.

EICHLER, M., DAHLHAUS, R. and DUECK, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *J. Time Series Anal.* **38** 225–242. MR3611742 https://doi.org/10.1111/jtsa.12213

EKBLOM, H. (1974). L_p-methods for robust regression. *BIT* **14** 22–32. MR0341759 https://doi.org/10.1007/bf01933114

ENGELKE, S. and HITZ, A. S. (2020). Graphical models for extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 871–932. With discussions. MR4136498

ENGELKE, S., HITZ, A. S. and GNECCO, N. (2019). graphicalExtremes: Statistical methodology for graphical extreme value models. R package version 0.1.0. Available at https://CRAN.R-project.org/package=graphicalExtremes.

FALLANI, F. D. V., CORAZZOL, M., STERNBERG, J. R., WYART, C. and CHAVEZ, M. (2015). Hierarchy of neural organization in the embryonic spinal cord: Granger-causality graph analysis of in vivo calcium imaging data. *IEEE Trans. Neural Syst. Rehabil. Eng.* **23** 333–341.

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

FRIEDRICH, J. and PANINSKI, L. (2016). Fast active set methods for online spike inference from calcium imaging. *Adv. Neural Inf. Process. Syst.* **29**.

GAO, X., SHEN, W., TING, C. M., CRAMER, S. C., SRINIVASAN, R. and OMBAO, H. (2018). Modeling brain connectivity with graphical models on frequency domain. arXiv preprint, arXiv:1810.03279.

GAROFALO, M., NIEUS, T., MASSOBRIO, P. and MARTINOIA, S. (2009). Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks. *PLoS ONE* **4** e6482.

GÓMEZ, E., GÓMEZ-VILLEGAS, M. A. and MARÍN, J. M. (1998). A multivariate generalization of the power exponential family of distributions. *Comm. Statist. Theory Methods* **27** 589–600. MR1619030 https://doi.org/10.1080/03610929808832115

GOODMAN, I. R. and KOTZ, S. (1973). Multivariate θ-generalized normal distributions. *J. Multivariate Anal.* **3** 204–219. MR0328996 https://doi.org/10.1016/0047-259X(73)90023-7

HAMMERSLEY, J. M. and CLIFFORD, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript, 46.

JEWELL, S. and WITTEN, D. (2018). Exact spike train inference via ℓ_0 optimization. *Ann. Appl. Stat.* **12** 2457–2482. MR3875708 https://doi.org/10.1214/18-AOAS1162

KEELEY, S. L., ZOLTOWSKI, D. M., AOI, M. C. and PILLOW, J. W. (2020). Modeling statistical dependencies in multi-region spike train data. *Curr. Opin. Neurobiol.* **65** 194–202. https://doi.org/10.1016/j.conb.2020.11.005

KRUMIN, M. and SHOHAM, S. (2010). Multivariate autoregressive modeling and Granger causality analysis of multiple spike trains. *Comput. Intell. Neurosci.* **2010** 752428. https://doi.org/10.1155/2010/752428

LAMBERT, R. C., TULEAU-MALOT, C., BESSAIH, T., RIVOIRARD, V., BOURET, Y., LERESCHE, N. and REYNAUD-BOURET, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *J. Neurosci. Methods* **297** 9–21. https://doi.org/10.1016/j.jneumeth.2017.12.026

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford Univ. Press, New York. Oxford Science Publications. MR1419991

LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electron. J. Stat.* **9** 608–642. MR3331852 https://doi.org/10.1214/15-EJS1013

LEIN, E. S., HAWRYLYCZ, M. J., AO, N., AYRES, M., BENSINGER, A., BERNARD, A., BOE, A. F., BOGUSKI, M. S., BROCKWAY, K. S. et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445** 168–176.

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst.* **24** 1432–1440.

MASUD, M. S. and BORISYUK, R. (2011). Statistical technique for analysing functional connectivity of multiple spike trains. *J. Neurosci. Methods* **196** 201–219.

MCINTOSH, A. R. and GONZALEZ-LIMA, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2** 2–22.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

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

MONEY, A. H., AFFLECK-GRAVES, J. F., HART, M. L. and BARR, G. D. I. (1982). The linear regression model: Lp norm estimation and the choice of p. *Comm. Statist. Simulation Comput.* **11** 89–109.

NADARAJAH, S. (2005). A generalized normal distribution. *J. Appl. Stat.* **32** 685–694. MR2119411 https://doi.org/10.1080/02664760500079464

NYQUIST, H. (1983). The optimal L_p norm estimator in linear regression models. *Comm. Statist. Theory Methods* **12** 2511–2524. MR0715180 https://doi.org/10.1080/03610928308828618

PARK, C. H., KIM, S. Y., KIM, Y. H. and KIM, K. (2008). Comparison of the small-world topology between anatomical and functional connectivity in the human brain. *Phys. A, Stat. Mech. Appl.* **387** 5958–5962.

PARK, I. M., ARCHER, E. W., PRIEBE, N. and PILLOW, J. W. (2013). Spectral methods for neural characterization using generalized quadratic models. *Adv. Neural Inf. Process. Syst.* **26**.

PNEVMATIKAKIS, E. A., SOUDRY, D., GAO, Y., MACHADO, T. A., MEREL, J., PFAU, D., REARDON, T., MU, Y., LACEFIELD, C. et al. (2016). Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89** 285–299.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2013). Inference of functional connectivity in neurosciences via Hawkes processes. In 2013 *IEEE Global Conference on Signal and Information Processing* 317–320.

ROCCO, M. (2014). Extreme value theory in finance: A survey. *J. Econ. Surv.* **28** 82–108.

ROSSANT, C., KADIR, S. N., GOODMAN, D. F., SCHULMAN, J., HUNTER, M. L., SALEEM, A. B., GROSMARK, A., BELLUSCIO, M., DENFIELD, G. H. et al. (2016). Spike sorting for large, dense electrode arrays. *PLoS Comput. Biol.* **11** e1004083.

SAKIA, K. and MIYASHITA, Y. (1994). Neuronal tuning to learned complex forms in vision. *NeuroReport* **5** 829–832.

SO, K., KORALEK, A. C., GANGULY, K., GASTPAR, M. C. and CARMENA, J. M. (2012). Assessing functional connectivity of neural ensembles using directed information. *J. Neural Eng.* **9** 026004.

SPORNS, O., HONEY, C. J. and KÖTTER, R. (2007). Identification and classification of hubs in brain networks. *PLoS ONE* **2** e1049.

STAM, C. J. (2004). Functional connectivity patterns of human magnetoencephalographic recordings: A 'small-world' network? *Neurosci. Lett.* **355** 25–28. https://doi.org/10.1016/j.neulet.2003.10.063

TALIH, M. and HENGARTNER, N. (2005). Structural learning with time-varying components: Tracking the cross-section of the financial time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 321–341. MR2155341 https://doi.org/10.1111/j.1467-9868.2005.00504.x

TIBAU, E., VALENCIA, M. and SORIANO, J. (2013). Identification of neuronal network properties from the spectral analysis of calcium imaging signals in neuronal cultures. *Front. Neural Circuits* **7** 199. https://doi.org/10.3389/fncir.2013.00199

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

VERDOOLAEGE, G. and SCHEUNDERS, P. (2012). On the geometry of multivariate generalized Gaussian models. *J. Math. Imaging Vision* **43** 180–193. MR2910882 https://doi.org/10.1007/s10851-011-0297-8

VOGELSTEIN, J. T., PACKER, A. M., MACHADO, T. A., SIPPY, T., BABADI, B., YUSTE, R. and PANINSKI, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104** 3691–3704.

WANG, T., REN, Z., DING, Y., FANG, Z., SUN, Z., MACDONALD, M. L., SWEET, R. A., WANG, J. and CHEN, W. (2016). FastGGM: An efficient algorithm for the inference of Gaussian graphical model in biological networks. *PLoS Comput. Biol.* **12** e1004755.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2012). Graphical models via generalized linear models. In *NeurIPS* **25** 1367–1375.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YATSENKO, D., JOSIĆ, K., ECKER, A. S., FROUDARAKIS, E., COTTON, R. J. and TOLIAS, A. S. (2015). Improved estimation and interpretation of correlations in neural circuits. *PLoS Comput. Biol.* **11** e1004083.

YU, H., CHOO, Z., UY, W. I. T., DAUWELS, J. and JONATHAN, P. (2012). Modeling extreme events in spatial domain by copula graphical models. In 2012 15*th International Conference on Information Fusion* 1761–1768.

ZAATOUR, R. (2014). hawkes: Hawkes process simulation and calibration toolkit. R package version 0.0-4. Available at https://CRAN.R-project.org/package=hawkes.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2387–2409
https://doi.org/10.1214/22-AOAS1724

DOUBLY-ONLINE CHANGEPOINT DETECTION FOR MONITORING HEALTH STATUS DURING SPORTS ACTIVITIES

BY MATTIA STIVAL[1,a], MAURO BERNARDI[1,b] AND PETROS DELLAPORTAS[2,3,c]

[1]*Department of Statistical Sciences, University of Padova,* [a]*mattia.stival@unipd.it,* [b]*mauro.bernardi@unipd.it*
[2]*Department of Statistical Science, University College London,* [c]*p.dellaportas@ucl.ac.uk*
[3]*Department of Statistics, Athens University of Economic and Business*

We provide an online framework for analyzing data recorded by smart watches during running activities. In particular, we focus on identifying variations in the behavior of one or more measurements caused by changes in physical condition, such as physical discomfort, periods of prolonged detraining, or even the malfunction of measuring devices. Our framework considers data as a sequence of running activities represented by multivariate time series of physical and biometric data. We combine classical changepoint detection models with an unknown number of components with Gaussian state space models to detect distributional changes between a sequence of activities. The model considers multiple sources of dependence due to the sequential nature of subsequent activities, the autocorrelation structure within each activity, and the contemporaneous dependence between different variables. We provide an online expectation-maximization (EM) algorithm involving a sequential Monte Carlo (SMC) approximation of changepoint predicted probabilities. As a byproduct of our model assumptions, our proposed approach processes sequences of multivariate time series in a doubly-online framework. While classical changepoint models detect changes between subsequent activities, the state space framework, coupled with the online EM algorithm, provides the additional benefit of estimating the real-time probability that a current activity is a changepoint.

REFERENCES

ADAMS, R. P. and MACKAY, D. J. (2007). Bayesian online changepoint detection. ArXiv preprint. Available at arXiv:0710.3742.

AMINIKHANGHAHI, S. and COOK, D. J. (2017). A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **51** 339–367. https://doi.org/10.1007/s10115-016-0987-z

BOURDON, P. C., CARDINALE, M., MURRAY, A., GASTIN, P., KELLMANN, M., VARLEY, M. C., GABBETT, T. J., COUTTS, A. J., BURGESS, D. J. et al. (2017). Monitoring athlete training loads: Consensus statement. *Int. J. Sports Physiol.* **12** 161–170.

BUCHHEIT, M. (2014). Monitoring training status with HR measures: Do all roads lead to Rome? *Front. Physiol.* **5** 73. https://doi.org/10.3389/fphys.2014.00073

CAPPÉ, O. and MOULINES, E. (2009). On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 593–613. MR2749909 https://doi.org/10.1111/j.1467-9868.2009.00698.x

CARON, F., DOUCET, A. and GOTTARDO, R. (2012). On-line changepoint detection and parameter estimation with application to genomic data. *Stat. Comput.* **22** 579–595. MR2865037 https://doi.org/10.1007/s11222-011-9248-x

CHIB, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86** 221–241. MR1649222 https://doi.org/10.1016/S0304-4076(97)00115-2

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537

DENEVI, G., STAMOS, D., CILIBERTO, C. and PONTIL, M. (2019). Online-within-online meta-learning. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds.) **32**. Curran Associates, Red Hook.

Key words and phrases. Real-time health monitoring, smart watches, online expectation maximization, sequential Monte Carlo.

DONG, J.-G. (2016). The role of heart rate variability in sports physiology. *Exp. Ther. Med.* **11** 1531–1536. https://doi.org/10.3892/etm.2016.3104

DURBIN, J. and KOOPMAN, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed. *Oxford Statistical Science Series* **38**. Oxford Univ. Press, Oxford. MR3014996 https://doi.org/10.1093/acprof:oso/9780199641178.001.0001

DU ROY DE CHAUMARAY, M., MARBAC, M. and NAVARRO, F. (2020). Mixture of hidden Markov models for accelerometer data. *Ann. Appl. Stat.* **14** 1834–1855. MR4194250 https://doi.org/10.1214/20-AOAS1375

ELLIOTT, R. J., FORD, J. J. and MOORE, J. B. (2002). On-line almost-sure parameter estimation for partially observed discrete-time linear systems with known noise characteristics. *Internat. J. Adapt. Control Signal Process.* **16** 435–453.

FEARNHEAD, P. and LIU, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 589–605. MR2370070 https://doi.org/10.1111/j.1467-9868.2007.00601.x

FREE, C., PHILLIPS, G., GALLI, L., WATSON, L., FELIX, L., EDWARDS, P., PATEL, V. and HAINES, A. (2013). The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: A systematic review. *PLoS Med.* **10** 1–45.

FRICK, H. and KOSMIDIS, I. (2017). trackeR: Infrastructure for running and cycling data from GPS-enabled tracking devices in R. *J. Stat. Softw.* **82** 1–29.

GARCÍA-GUZMÁN, J. J., PÉREZ-RÀFOLS, C., CUARTERO, M. and CRESPO, G. A. (2021). Microneedle based electrochemical (Bio)Sensing: Towards decentralized and continuous health status monitoring. *TrAC, Trends Anal. Chem.* **135** 116148.

GRUNDY, T., KILLICK, R. and MIHAYLOV, G. (2020). High-dimensional changepoint detection via a geometrically inspired mapping. *Stat. Comput.* **30** 1155–1166. MR4108696 https://doi.org/10.1007/s11222-020-09940-y

HAYNES, K., FEARNHEAD, P. and ECKLEY, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Stat. Comput.* **27** 1293–1305. MR3647098 https://doi.org/10.1007/s11222-016-9687-5

HUANG, L., BAI, J., IVANESCU, A., HARRIS, T., MAURER, M., GREEN, P. and ZIPUNNIKOV, V. (2019). Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *J. Amer. Statist. Assoc.* **114** 553–564. MR3963162 https://doi.org/10.1080/01621459.2018.1482750

JUNGBACKER, B. and KOOPMAN, S. J. (2008). Likelihood-based analysis for dynamic factor models. Technical Report, Tinbergen Institute Discussion Paper.

KANTAS, N., DOUCET, A., SINGH, S. S., MACIEJOWSKI, J. and CHOPIN, N. (2015). On particle methods for parameter estimation in state-space models. *Statist. Sci.* **30** 328–351. MR3383884 https://doi.org/10.1214/14-STS511

KITAGAWA, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Statist. Assoc.* **82** 1032–1041. MR0922169

LUO, L. and SONG, P. X.-K. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 69–97. MR4060977

PELLICCIA, A., SHARMA, S., GATI, S., BÄCK, M., BÖRJESSON, M., CASELLI, S., COLLET, J.-P., CORRADO, D., DREZNER, J. A. et al. (2021). 2020 ESC Guidelines on sports cardiology and exercise in patients with cardiovascular disease: The Task Force on sports cardiology and exercise in patients with cardiovascular disease of the European Society of Cardiology (ESC). *Eur. Heart J.* **42** 17–96.

PKVITALITY (2020). *PKvitality faqs*. Available at https://www.pkvitality.com/.

QIAN, T., YOO, H., KLASNJA, P., ALMIRALL, D. and MURPHY, S. A. (2021). Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika* **108** 507–527. MR4298759 https://doi.org/10.1093/biomet/asaa070

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SAMÉ, A. and GOVAERT, G. (2017). Segmental dynamic factor analysis for time series of curves. *Stat. Comput.* **27** 1617–1637. MR3687329 https://doi.org/10.1007/s11222-016-9707-5

SCHNEIDER, C., HANAKAM, F., WIEWELHOVE, T., DÖWELING, A., KELLMANN, M., MEYER, T., PFEIFFER, M. and FERRAUTI, A. (2018). Heart rate monitoring in team sports-a conceptual framework for contextualizing heart rate measures for training and recovery prescription. *Front. Physiol.* **9** 639. https://doi.org/10.3389/fphys.2018.00639

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

SINGH, N., MONEGHETTI, K. J., CHRISTLE, J. W., HADLEY, D., FROELICHER, V. and PLEWS, D. (2018). Heart rate variability: An old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. Part two: Prognosis and training. *Arrhythm Electrophysiol. Rev.* **7** 247–255. https://doi.org/10.15420/aer.2018.30.2

SIQUEIRA DO PRADO, L. S., CARPENTIER, C., PREAU, M., SCHOTT, A.-M. and DIMA, A. L. (2019). Behavior change content, understandability, and actionability of chronic condition self-management apps available in France: Systematic search and evaluation. *JMIR mHealth uHealth* **7** e13494. https://doi.org/10.2196/13494

SONG, P. X.-K. (2007). *Correlated Data Analysis*: *Modeling*, *Analytics*, *and Applications*. *Springer Series in Statistics*. Springer, New York. MR2377853

SOUISSI, A., HADDAD, M., DERGAA, I., SAAD, H. B. and CHAMARI, K. (2021). A new perspective on cardiovascular drift during prolonged exercise. *Life Sci*. **287** 120109. https://doi.org/10.1016/j.lfs.2021.120109

STATISTA (2020a). *Running and Jogging—Statistics and Facts*. Available at https://www.statista.com/topics/1743/running-and-jogging/#dossierSummary.

STATISTA (2020b). *Wearables dossier*. Available at https://www.statista.com/study/15607/wearables-statista-dossier/.

STATISTA (2020c). *eServices Report 2020—Fitness*. Available at https://www.statista.com/study/36674/fitness-report/.

STIVAL, M., BERNARDI, M. and DELLAPORTAS, P. (2023). Supplement to "Doubly-online changepoint detection for monitoring health status during sports activities." https://doi.org/10.1214/22-AOAS1724SUPPA, https://doi.org/10.1214/22-AOAS1724SUPPB

TITSIAS, M. K., SYGNOWSKI, J. and CHEN, Y. (2022). Sequential changepoint detection in neural networks with checkpoints. *Stat*. *Comput*. **32** Paper No. 26. MR4394856 https://doi.org/10.1007/s11222-022-10088-0

VITABILE, S., MARKS, M., STOJANOVIC, D., PLLANA, S., MOLINA, J. M., KRZYSZTON, M., SIKORA, A., JARYNOWSKI, A., HOSSEINPOUR, F. et al. (2019). Medical Data Processing and Analysis for Remote Health and Activities Monitoring. In *High-Performance Modelling and Simulation for Big Data Applications*: *Selected Results of the COST Action IC*1406 *CHiPSet* 186–220. Springer, Cham.

XIE, L., ZOU, S., XIE, Y. and VEERAVALLI, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *JSAIT* **2** 494–514.

YILDIRIM, S., SINGH, S. S. and DOUCET, A. (2013). An online expectation-maximization algorithm for changepoint models. *J*. *Comput*. *Graph*. *Statist*. **22** 906–926. MR3173749 https://doi.org/10.1080/10618600.2012.674653

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2410–2431
https://doi.org/10.1214/22-AOAS1725

SIGNAL-NOISE RATIO OF GENETIC ASSOCIATIONS AND STATISTICAL POWER OF SNP-SET TESTS

BY HONG ZHANG[1,a], MING LIU[2,b], JIASHUN JIN[3,d] AND ZHEYANG WU[2,c]

[1]*Translational Biomarker Statistics, Global Biometrics and Data Management, Pfizer Inc.,* [a]*hong.zhang3@pfizer.com*
[2]*Department of Mathematical Sciences, Worcester Polytechnic Institute,* [b]*mliu5@wpi.edu,* [c]*zheyangwu@wpi.edu*
[3]*Department of Statistics, Carnegie Mellon University,* [d]*jiashun@cmu.edu*

The SNP-set analysis is a powerful tool for dissecting the genetics of complex human diseases. There are three fundamental genetic association approaches to SNR-set analysis: the marginal model fitting approach, the joint model fitting approach, and the decorrelation approach. A problem of primary interest is how these approaches compare with each other. To address this problem, we develop a theoretical platform to compare the signal-to-noise ratio (SNR) of these approaches under the generalized linear model. We elaborate how causal genetic effects give rise to statistically detectable association signals and show that, when causal effects spread over blocks of strong linkage disequilibrium (LD), the SNR of the marginal model fitting is usually higher than that of the decorrelation approach which, in turn, is higher than that of the unbiased joint model fitting approach. We also scrutinize dense effects and LDs by a bivariate model and extensive simulations using the 1000 Genome Project data. Last, we compare the statistical power of two generic types of SNP-set tests (summation-based and supremum-based) by simulations and an osteoporosis study using large data from UK Biobank. Our results help develop powerful tools for SNP-set analysis and understand the signal detection problem in the presence of colored noise.

REFERENCES

ABRAHAM, G., QIU, Y. and INOUYE, M. (2017). FlashPCA2: Principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33** 2776–2778. https://doi.org/10.1093/bioinformatics/btx299

ARIAS-CASTRO, E., HUANG, R. and VERZELEN, N. (2020). Detection of sparse positive dependence. *Electron. J. Stat.* **14** 702–730. MR4057605 https://doi.org/10.1214/19-EJS1675

ARIAS-CASTRO, E. and WANG, M. (2017). Distribution-free tests for sparse heterogeneous mixtures. *TEST* **26** 71–94. MR3613606 https://doi.org/10.1007/s11749-016-0499-x

BARNETT, I., MUKHERJEE, R. and LIN, X. (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *J. Amer. Statist. Assoc.* **112** 64–76. MR3646553 https://doi.org/10.1080/01621459.2016.1192039

DEY, R., SCHMIDT, E. M., ABECASIS, G. R. and LEE, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101** 37–49.

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. L. and KIPNIS, A. (2021). The impossibility region for detecting sparse mixtures using the higher criticism. ArXiv Preprint. Available at arXiv:2103.03218.

EVANGELOU, E., KERKHOF, H. J., STYRKARSDOTTIR, U., NTZANI, E. E., BOS, S. D., ESKO, T., EVANS, D. S., METRUSTRY, S., PANOUTSOPOULOU, K. et al. (2014). A meta-analysis of genome-wide association studies identifies novel variants associated with osteoarthritis of the hip. *Ann. Rheum. Dis.* **73** 2130–2136.

FAHRMEIR, L. (1987). Asymptotic testing theory for generalized linear models. *Statistics* **18** 65–76. MR0871451 https://doi.org/10.1080/02331888708801992

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368. MR0773172 https://doi.org/10.1214/aos/1176346597

Key words and phrases. SNP-set analysis, causal genetic effect, linkage disequilibrium, signal-noise ratio, global hypothesis test, osteoporosis.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861 https://doi.org/10.1214/10-AOS798
FISHER, R. A. (1934). *Statistical Methods for Research Workers*, 5th ed. Oliver and Boyd, Edinburgh.
GUO, B. and WU, B. (2019). Powerful and efficient SNP-set association tests across multiple phenotypes using GWAS summary data. *Bioinformatics* **35** 1366–1372.
GUO, L., HAN, J., GUO, H., LV, D. and WANG, Y. (2019). Pathway and network analysis of genes related to osteoporosis. *Mol. Med. Rep.* **20** 985–994.
HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764
HE, S. and WU, Z. (2011). Gene-based Higher Criticism methods for large-scale exonic single-nucleotide polymorphism data. In *BMC Proceedings* **5** S65. Springer, Berlin.
HOH, J., WILLE, A. and OTT, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* **11** 2115–2119. https://doi.org/10.1101/gr.204001
HOTELLING, H. (1931). The generalization of student's ratio. *Ann. Math. Stat.* **2** 360–378.
KE, Z. T., FAN, J. and WU, Y. (2015). Homogeneity pursuit. *J. Amer. Statist. Assoc.* **110** 175–194. MR3338495 https://doi.org/10.1080/01621459.2014.892882
KOVAC, M., WOOLLEY, C., RIBI, S., BLATTMANN, C., ROTH, E., MORINI, M., KOVACOVA, M., AMELINE, B., KULOZIK, A. et al. (2021). Germline RET variants underlie a subset of paediatric osteosarcoma. *J. Med. Genet.* **58** 20–24.
KWAK, I.-Y. and PAN, W. (2016). Adaptive gene- and pathway-trait association testing with GWAS summary statistics. *Bioinformatics* **32** 1178–1184. https://doi.org/10.1093/bioinformatics/btv719
LEHNER, B., SEMPLE, J. I., BROWN, S. E., COUNSELL, D., CAMPBELL, R. D. and SANDERSON, C. M. (2004). Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics* **83** 153–167.
LITTELL, R. C. and FOLKS, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests. II. *J. Amer. Statist. Assoc.* **68** 193–194. MR0375577
LUO, L., PENG, G., ZHU, Y., DONG, H., AMOS, C. I. and XIONG, M. (2010). Genome-wide gene and pathway analysis. *Eur. J. Hum. Genet.* **18** 1045–1053.
MARCHINI, J., DONNELLY, P. and CARDON, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417. https://doi.org/10.1038/ng1537
MASCOLO, E., LIGUORI, F., STUFERA MECARELLI, L., AMOROSO, N., MERIGLIANO, C., AMADIO, S., VOLONTÉ, C., CONTESTABILE, R., TRAMONTI, A. et al. (2021). Functional inactivation of drosophila GCK orthologs causes genomic instability and oxidative stress in a fly model of MODY-2. *Int. J. Mol. Sci.* **22** 918.
MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. CRC Press LLC, FL.
MORRIS, J. A., KEMP, J. P., YOULTEN, S. E., LAURENT, L., LOGAN, J. G., CHAI, R. C., VULPESCU, N. A., FORGETTA, V., KLEINMAN, A. et al. (2019). An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51** 258–266.
PASANIUC, B. and PRICE, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18** 117–127. https://doi.org/10.1038/nrg.2016.142
PUN, K. K., LAU, P. and HO, P. W. (1989). The characterization, regulation, and function of insulin receptors on osteoblast-like clonal osteosarcoma cell line. *J. Bone Miner. Res.* **4** 853–862. https://doi.org/10.1002/jbmr.5650040610
QI, X., YU, X.-J., WANG, X.-M., SONG, T.-N., ZHANG, J., GUO, X.-Z., LI, G.-J. and SHAO, M. (2019). Knockdown of KCNQ1OT1 suppresses cell invasion and sensitizes osteosarcoma cells to CDDP by upregulating DNMT1-mediated Kcnq1 expression. *Mol. Ther. Nucleic Acids* **17** 804–818.
ROSA, S., RUFINO, A., JUDAS, F., TENREIRO, C., LOPES, M. and MENDES, A. (2011). Expression and function of the insulin receptor in normal and osteoarthritic human chondrocytes: Modulation of anabolic gene expression, glucose transport and GLUT-1 content by insulin. *Osteoarthr. Cartil.* **19** 719–727.
SCHORK, N. J., MURRAY, S. S., FRAZER, K. A. and TOPOL, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Option Genet. Dev.* **19** 212–219.
SHAO, J. (2010). *Mathematical Statistics*, 2nd ed. Springer, Berlin.
SIVA, N. (2008). 1000 genomes project. *Nat. Biotechnol.* **26** 256–256.
STELZER, G., ROSEN, N., PLASCHKES, I., ZIMMERMAN, S., TWIK, M., FISHILEVICH, S., STEIN, T. I., NUDEL, R., LIEDER, I. et al. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54** 1–30.
STOUFFER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS, R. M. (1949). *The American Soldier*: *Adjustment During Army Life* **I**. Princeton Univ. Press, NJ.
SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J. et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12** e1001779.

WU, Z. and ZHAO, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet*. **5**.

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

WU, Z., SUN, Y., HE, S., CHO, J., ZHAO, H. and JIN, J. (2014). Detection boundary and higher criticism approach for rare and weak genetic effects. *Ann. Appl. Stat*. **8** 824–851. MR3262536 https://doi.org/10.1214/14-AOAS724

YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., WEEDON, M. N. et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet*. **44** 369.

ZENG, W., LIU, Q., CHEN, Z., WU, X., ZHONG, Y. and WU, J. (2016). Silencing of hERG1 gene inhibits proliferation and invasion, and induces apoptosis in human osteosarcoma cells by targeting the NF-κB pathway. *J. Cancer* **7** 746.

ZHANG, H., JIN, J. and WU, Z. (2020). Distributions and power of optimal signal-detection statistics in finite case. *IEEE Trans. Signal Process*. **68** 1021–1033. MR4114605 https://doi.org/10.1109/TSP.2020.2967179

ZHANG, H. and WU, Z. (2022). The general goodness-of-fit tests for correlated data. *Comput. Statist. Data Anal*. **167** 107379. MR4333746 https://doi.org/10.1016/j.csda.2021.107379

ZHANG, H., TONG, T., LANDERS, J. and WU, Z. (2020). TFisher: A powerful truncation and weighting procedure for combining p-values. *Ann. Appl. Stat*. **14** 178–201. MR4085089 https://doi.org/10.1214/19-AOAS1302

ZHANG, H., LIU, M., JIN, J. and WU, Z. (2023). Supplement to "On signal-noise ratio of causal genetic effects and statistical power of SNP-set tests." https://doi.org/10.1214/22-AOAS1725SUPPA, https://doi.org/10.1214/22-AOAS1725SUPPB, https://doi.org/10.1214/22-AOAS1725SUPPC

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2432–2451
https://doi.org/10.1214/22-AOAS1726

EVALUATING THE USE OF GENERALIZED DYNAMIC WEIGHTED ORDINARY LEAST SQUARES FOR INDIVIDUALIZED HIV TREATMENT STRATEGIES

BY LARRY DONG[1,a], ERICA E. M. MOODIE[1,b], LAURA VILLAIN[2,c] AND RODOLPHE THIÉBAUT[2,d]

[1]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University,* [a]*larry.dong@mail.mcgill.ca,* [b]*erica.moodie@mcgill.ca*

[2]*University of Bordeaux, INSERM U1219 Bordeaux Population Health, Inria SISTM,* [c]*laura.villain@u-bordeaux.fr,* [d]*rodolphe.thiebaut@u-bordeaux.fr*

A dynamic treatment regimes (DTR) represents a statistical paradigm in precision medicine which aims to optimize patient outcomes by individualizing treatments. At its simplest, a DTR may require only a single decision to be made; this special case is called an individualized treatment rule (ITR) and is often used to maximize short-term rewards. Generalized dynamic weighted ordinary least squares (G-dWOLS), a DTR estimation method that offers theoretical advantages such as double robustness of parameter estimators in the decision rules, has been recently extended to accommodate categorical treatments. In this work G-dWOLS is applied to *longitudinal* data to estimate an optimal ITR. This novel method is demonstrated in simulations and is then applied to a population affected by HIV, whereby an ITR for the administration of Interleukin 7 (IL-7) is devised to maximize the duration where the CD4 load is above a healthy threshold (500 cells/μL) while preventing the administration of unnecessary injections.

REFERENCES

AUSTIN, P. C. (2018). Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes. *Stat. Med.* **37** 1874–1894. MR3799846 https://doi.org/10.1002/sim.7615

AUSTIN, P. C. and STUART, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140 https://doi.org/10.1002/sim.6607

CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes. Statistics for Biology and Health*. Springer, New York. MR3112454 https://doi.org/10.1007/978-1-4614-7428-9

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

DONG, L., MOODIE, E. E., VILLAIN, L. and THIÉBAUT, R. (2023). Supplement to "Evaluating the use of generalized dynamic weighted ordinary least squares for individualized HIV treatment strategies." https://doi.org/10.1214/22-AOAS1726SUPP

DOUEK, D. C., ROEDERER, M. and KOUP, R. A. (2009). Emerging concepts in the immunopathogenesis of AIDS. *Annu. Rev. Med.* **60** 471–484. https://doi.org/10.1146/annurev.med.60.041807.123549

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

FLURY, B. K. and RIEDWYL, H. (1986). Standard distance in univariate and multivariate analysis. *Amer. Statist.* **40** 249–251.

GELBER, R. D., GOLDHIRSCH, A. and CAVALLI, F. (1991). Quality-of-life-adjusted evaluation of adjuvant therapies for operable breast cancer. The international breast cancer study group. *Ann. Intern. Med.* **114** 621–628. https://doi.org/10.7326/0003-4819-114-8-621

GELBER, R. D., COLE, B. F., GELBER, S. and GOLDHIRSCH, A. (1995). Comparing treatments using quality-adjusted survival: The Q-TWiST method. *Amer. Statist.* **49** 161–169.

Key words and phrases. Dynamic treatment regime, adaptive treatment strategy, precision medicine, individualized treatment rule, longitudinal data, HIV.

GLASZIOU, P. P., COLE, B. F., GELBER, R. D., HILDEN, J. and SIMES, R. J. (1998). Quality adjusted survival analysis with repeated quality of life measures. *Stat. Med.* **17** 1215–1229. https://doi.org/10.1002/(sici)1097-0258(19980615)17:11<1215::aid-sim844>3.0.co;2-y

GRABAR, S., LE MOING, V., GOUJARD, C., LEPORT, C., KAZATCHKINE, M. D., COSTAGLIOLA, D. and WEISS, L. (2000). Clinical outcome of patients with HIV-1 infection according to immunologic and virologic response after 6 months of highly active antiretroviral therapy. *Ann. Intern. Med.* **133** 401–410.

JARNE, A., COMMENGES, D., VILLAIN, L., PRAGUE, M., LÉVY, Y. and THIÉBAUT, R. (2017). Modeling $CD4^+$ T cells dynamics in HIV-infected patients receiving repeated cycles of exogenous Interleukin 7^1. *Ann. Appl. Stat.* **11** 1593–1616. MR3709571 https://doi.org/10.1214/17-AOAS1047

KLEINBERG, J. and TARDOS, E. (2006). *Algorithm Design*. Pearson Education, India.

KOSOROK, M. R. and MOODIE, E. E. (2015). *Adaptive Treatment Strategies in Practice*: *Planning Trials and Analyzing Data for Personalized Medicine* **21**. SIAM, Pennsylvania, USA.

LAWSON, B. R., GONZALEZ-QUINTIAL, R., ELEFTHERIADIS, T., FARRAR, M. A., MILLER, S. D., SAUER, K., MCGAVERN, D. B., KONO, D. H., BACCALA, R. et al. (2015). Interleukin-7 is required for CD4(+) T cell activation and autoimmune neuroinflammation. *Clin. Immunol.* **161** 260–269. https://doi.org/10.1016/j.clim.2015.08.007

LEVY, Y., LACABARATZ, C., WEISS, L., VIARD, J.-P., GOUJARD, C., LELIÈVRE, J.-D., BOUÉ, F., MOLINA, J.-M., ROUZIOUX, C. et al. (2009). Enhanced T cell recovery in HIV-1–infected adults through IL-7 treatment. *J. Clin. Invest.* **119** 997–1007.

LEVY, Y., SERETI, I., TAMBUSSI, G., ROUTY, J., LELIEVRE, J., DELFRAISSY, J., MOLINA, J., FISCHL, M., GOUJARD, C. et al. (2012). Effects of recombinant human Interleukin 7 on T-cell recovery and thymic output in HIV-infected patients receiving antiretroviral therapy: Results of a phase I/IIa randomized, placebo-controlled, multicenter study. *Clin. Infect. Dis.* **55** 291–300.

LEWDEN, C., CHÊNE, G., MORLAT, P., RAFFI, F., DUPON, M., DELLAMONICA, P., PELLEGRIN, J.-L., KATLAMA, C., DABIS, F. et al. (2007). HIV-infected adults with a CD4 cell count greater than 500 cells/mm3 on long-term combination antiretroviral therapy reach same mortality rates as the general population. *J. Acquir. Immune Defic. Syndr.* **46** 72–77.

LIU, N., LIU, Y., LOGAN, B., XU, Z., TANG, J. and WANG, Y. (2019). Learning the dynamic treatment regimes from medical registry data through deep Q-network. *Sci. Rep.* **9** 1–10.

LOGEROT, S., RANCEZ, M., MUYLDER, B. C., FIGUEIREDO-MORGADO, S., ROZLAN, S., TAMBUSSI, G., BEQ, S., COUËDEL-COURTEILLE, A. and CHEYNIER, R. (2018). HIV reservoir dynamics in HAART-treated poor immunological responder patients under IL-7 therapy. *AIDS* **32** 715–720. https://doi.org/10.1097/QAD.0000000000001752

MACKALL, C. L., FRY, T. J. and GRESS, R. E. (2011). Harnessing the biology of IL-7 for therapeutic application. *Nat. Rev., Immunol.* **11** 330–342. https://doi.org/10.1038/nri2970

MOODIE, E. E. M., CHAKRABORTY, B. and KRAMER, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canad. J. Statist.* **40** 629–645. MR2998853 https://doi.org/10.1002/cjs.11162

MURPHY, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Stat. Med.* **24** 1455–1481. MR2137651 https://doi.org/10.1002/sim.2022

MURPHY, S. A. (2005b). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

OPPORTUNISTIC INFECTIONS PROJECT TEAM OF THE COLLABORATION OF OBSERVATIONAL HIV EPIDEMIOLOGICAL RESEARCH IN EUROPE (COHERE) IN EUROCOORD (2012). CD4 cell count and the risk of AIDS or death in HIV-infected adults on combination antiretroviral therapy with a suppressed viral load: A longitudinal cohort study from COHERE. *PLoS Med.* 9.

PAPADOGEORGOU, G., CHOIRAT, C. and ZIGLER, C. M. (2019). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics* **20** 256–272. MR3922132 https://doi.org/10.1093/biostatistics/kxx074

PASIN, C., DUFOUR, F., VILLAIN, L., ZHANG, H. and THIÉBAUT, R. (2018). Controlling IL-7 injections in HIV-infected patients. *Bull. Math. Biol.* **80** 2349–2377. MR3844626 https://doi.org/10.1007/s11538-018-0465-8

PETERSEN, M. L., DEEKS, S. G. and VAN DER LAAN, M. J. (2007). Individualized treatment rules: Generating candidate clinical trials. *Stat. Med.* **26** 4578–4601. MR2411889 https://doi.org/10.1002/sim.2888

PRAGUE, M., COMMENGES, D., DRYLEWICZ, J. and THIÉBAUT, R. (2012). Treatment monitoring of HIV-infected patients based on mechanistic models. *Biometrics* **68** 902–911. MR3055195 https://doi.org/10.1111/j.1541-0420.2012.01749.x

QI, Z., LIU, D., FU, H. and LIU, Y. (2020). Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *J. Amer. Statist. Assoc.* **115** 678–691. MR4107672 https://doi.org/10.1080/01621459.2018.1529597

RAMASWAMI, R., BAYER, R. and GALEA, S. (2018). Precision medicine from a public health perspective. *Annu. Rev. Public Health* **39** 153–168. https://doi.org/10.1146/annurev-publhealth-040617-014158

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11

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

ROSENBERG, S. A., SPORTÈS, C., AHMADZADEH, M., FRY, T. J., NGO, L. T., SCHWARZ, S. L., STETLER-STEVENSON, M., MORTON, K. E., MAVROUKAKIS, S. A. et al. (2006). IL-7 administration to humans leads to expansion of CD8+ and CD4+ cells but a relative decrease of CD4+ T-regulatory cells. *J. Immunother.* **29** 313.

SCHULZ, J. and MOODIE, E. E. M. (2021). Doubly robust estimation of optimal dosing strategies. *J. Amer. Statist. Assoc.* **116** 256–268. MR4227692 https://doi.org/10.1080/01621459.2020.1753521

SIMONEAU, G., MOODIE, E. E. M., NIJJAR, J. S., PLATT, R. W. and THE SCOTTISH EARLY RHEUMATOID ARTHRITIS INCEPTION COHORT INVESTIGATORS (2020). Estimating optimal dynamic treatment regimes with survival outcomes. *J. Amer. Statist. Assoc.* **115** 1531–1539. MR4143483 https://doi.org/10.1080/01621459.2019.1629939

SPORTÈS, C., HAKIM, F. T., MEMON, S. A., ZHANG, H., CHUA, K. S., BROWN, M. R., FLEISHER, T. A., KRUMLAUF, M. C., BABB, R. R. et al. (2008). Administration of rhIL-7 in humans increases in vivo TCR repertoire diversity by preferential expansion of naive T cell subsets. *J. Exp. Med.* **205** 1701–1714.

STUART, E. A., LEE, B. K. and LEACY, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* **66** S84–S90.

SULLIVAN PEPE, M. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951.

SURH, C. D. and SPRENT, J. (2005). Regulation of mature T cell homeostasis. *Semin. Immunol.* **17** 183–191. https://doi.org/10.1016/j.smim.2005.02.007

SURH, C. D. and SPRENT, J. (2008). Homeostasis of naive and memory T cells. *Immunity* **29** 848–862. https://doi.org/10.1016/j.immuni.2008.11.002

TAO, Y., WANG, L. and ALMIRALL, D. (2018). Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *Ann. Appl. Stat.* **12** 1914–1938. MR3852703 https://doi.org/10.1214/18-AOAS1137

THALL, P. F., NGUYEN, H. Q., BRAUN, T. M. and QAZILBASH, M. H. (2013). Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics* **69** 673–682. MR3106595 https://doi.org/10.1111/biom.12065

THIÉBAUT, R., DRYLEWICZ, J., PRAGUE, M., LACABARATZ, C., BEQ, S., JARNE, A., CROUGHS, T., SEKALY, R.-P., LEDERMAN, M. M. et al. (2014). Quantifying and predicting the effect of exogenous Interleukin-7 on CD4+ T cells in HIV-1 infection. *PLoS Comput. Biol.* **10** e1003630.

THIÉBAUT, R., JARNE, A., ROUTY, J.-P., SERETI, I., FISCHL, M., IVE, P., SPECK, R. F., D'OFFIZI, G., CASARI, S. et al. (2016). Repeated cycles of recombinant human Interleukin 7 in HIV-infected patients with low CD4 T-cell reconstitution on antiretroviral therapy: Results of 2 phase II multicenter studies. *Clin. Infect. Dis.* **62** 1178–1185.

VILLAIN, L., COMMENGES, D., PASIN, C., PRAGUE, M. and THIÉBAUT, R. (2019). Adaptive protocols based on predictions from a mechanistic model of the effect of IL7 on CD4 counts. *Stat. Med.* **38** 221–235. MR3892816 https://doi.org/10.1002/sim.7957

WALLACE, M. P. and MOODIE, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71** 636–644. MR3402599 https://doi.org/10.1111/biom.12306

XUE, F., ZHANG, Y., ZHOU, W., FU, H. and QU, A. (2022). Multicategory angle-based learning for estimating optimal dynamic treatment regimes with censored data. *J. Amer. Statist. Assoc.* **117** 1438–1451. MR4480723 https://doi.org/10.1080/01621459.2020.1862671

YANG, D. and DALTON, J. E. (2012). A unified approach to measuring the effect size between two groups using SAS. In *SAS Global Forum* **335** 1–6.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2452–2472
https://doi.org/10.1214/22-AOAS1727

IMPUTATION SCORES

BY JEFFREY NÄF[a], META-LINA SPOHN[b], LORIS MICHEL[c] AND NICOLAI MEINSHAUSEN[d]

Seminar for Statistics, ETH Zürich, [a]*naef@stat.math.ethz.ch,* [b]*spohn@stat.math.ethz.ch,* [c]*loris.michel@gmail.com,* [d]*meinshausen@stat.math.ethz.ch*

Given the prevalence of missing data in modern statistical research, a broad range of methods is available for any given imputation task. How does one choose the "best" imputation method in a given application? The standard approach is to select some observations, set their status to missing, and compare prediction accuracy of the methods under consideration of these observations. Besides having to somewhat artificially mask observations, a shortcoming of this approach is that imputations based on the conditional mean will rank highest if predictive accuracy is measured with quadratic loss. In contrast, we want to rank highest an imputation that can sample from the true conditional distributions. In this paper we develop a framework called "Imputation Scores" (I-Scores) for assessing missing value imputations. We provide a specific I-Score, based on density ratios and projections, that is applicable to discrete and continuous data. It does not require to mask additional observations for evaluations and is also applicable if there are no complete observations. The population version is shown to be proper in the sense that the highest rank is assigned to an imputation method that samples from the correct conditional distribution. The propriety is shown under the *missing completely at random* (MCAR) assumption but is also shown to be valid under *missing at random* (MAR) with slightly more restrictive assumptions. We show empirically on a range of data sets and imputation methods that our score consistently ranks true data high(est) and is able to avoid pitfalls usually associated with performance measures such as RMSE. Finally, we provide the R-package Iscores available on CRAN with an implementation of our method.

REFERENCES

AUDIGIER, V., HUSSON, F. and JOSSE, J. (2016). Multiple imputation for continuous variables using a Bayesian principal component analysis. *J. Stat. Comput. Simul.* **86** 2140–2156. MR3491013 https://doi.org/10.1080/00949655.2015.1104683

BEAULAC, C. and ROSENTHAL, J. S. (2020). BEST: A decision tree algorithm that handles missing values. *Comput. Statist.* **35** 1001–1026. MR4133107 https://doi.org/10.1007/s00180-020-00987-z

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

BREIMAN, L. (2003b). Manual—setting up, using and understanding random forests v4.0.

BREIMAN, L. (2003a). Manual on setting up, using, and understanding Random Forests Technical report, Berkeley CA.

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

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. MR1926165 https://doi.org/10.1214/aos/1031689014

CAI, H., GOGGIN, B. and JIANG, Q. (2020). Two-sample test based on classification probability. *Stat. Anal. Data Min.* **13** 5–13. MR4063880 https://doi.org/10.1002/sam.11438

CHOI, J., DEKKERS, O. M. and LE CESSIE, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur. J. Epidemiol.* **34** 23–36. https://doi.org/10.1007/s10654-018-0447-z

Key words and phrases. Ranking, random projections, tree ensembles, random forest, KL-divergence.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345
DENG, Y., CHANG, C., IDO, M. S. and LONG, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci. Rep.* **6** 21689. https://doi.org/10.1038/srep21689
DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5
DING, Y. and ROSS, A. (2012). A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recognit.* **45** 919–933.
DOOVE, L. L., VAN BUUREN, S. and DUSSELDORP, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Statist. Data Anal.* **72** 92–104. MR3139350 https://doi.org/10.1016/j.csda.2013.10.025
GAFFERT, P., MEINFELDER, F. and BOSCH, V. (2016). Towards an MI-proper predictive mean matching. Discussion Paper.
GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988 https://doi.org/10.1198/jasa.2011.r10138
GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437
GNEITING, T., STANBERRY, L. I., GRIMIT, E. P., HELD, L. and JOHNSON, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* **17** 211–235. MR2434318 https://doi.org/10.1007/s11749-008-0114-x
GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Proceedings of the* 27*th International Conference on Neural Information Processing Systems – Volume* 2, *NIPS'*14 2672–2680. MIT Press, Cambridge, MA, USA.
HORTON, N. J. and LIPSITZ, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *Amer. Statist.* **55** 244–254. MR1963401 https://doi.org/10.1198/000313001317098266
JOSSE, J. and HUSSON, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70** 1–31.
JOSSE, J., PAGÈS, J. and HUSSON, F. (2011). Multiple imputation in principal component analysis. *Adv. Data Anal. Classif.* **5** 231–246. MR2832901 https://doi.org/10.1007/s11634-011-0086-7
LEISCH, F. and DIMITRIADOU, E. (2021). mlbench: Machine learning benchmark problems. R package version 2.1-3.
LI, J. and YU, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika* **80** 707–726. MR3392026 https://doi.org/10.1007/s11336-014-9410-4
LIN, W. and TSAI, C.-F. (2019). Missing value imputation: A review and analysis of the literature (2006–2017). *Artif. Intell. Rev.* **53** 1487–1509.
LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.
LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0890519
LUENGO, J., GARCÍA, S. and HERRERA, F. (2011). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowl. Inf. Syst.* **32** 1–32.
MALLEY, J. D., KRUPPA, J., DASGUPTA, A., MALLEY, K. G. and ZIEGLER, A. (2012). Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods Inf. Med.* **51** 74–81. https://doi.org/10.3414/ME00-01-0052
MAYER, I., SPORTISSE, A., JOSSE, J., TIERNEY, N. and VIALANEIX, N. (2021). R-miss-tastic: A unified platform for missing values methods and workflows. ArXiv preprint. Available at arXiv:1908.04822.
MEINSHAUSEN, N. (2010). Node harvest. *Ann. Appl. Stat.* **4** 2049–2072. MR2829946 https://doi.org/10.1214/10-AOAS367
MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558.
MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. MR3601702 https://doi.org/10.1080/01621459.2016.1174132
MUZELLEC, B., JOSSE, J., BOYER, C. and CUTURI, M. (2020). Missing data imputation using optimal transport. In *Proceedings of the* 37*th International Conference on Machine Learning* **119** 7130–7140. PMLR.
NÄF, J., SPOHN, M.-L., MICHEL, L. and MEINSHAUSEN, N. (2023). Supplement to "Imputation Scores." https://doi.org/10.1214/22-AOAS1727SUPPA, https://doi.org/10.1214/22-AOAS1727SUPPB
QUINLAN, M. (1993). *C*4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.
RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

SAAR-TSECHANSKY, M. and PROVOST, F. (2007). Handling missing values when applying classification models. *J. Mach. Learn. Res.* **8** 1623–1657.

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409 https://doi.org/10.1214/13-sts415

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

SHI, T. and HORVATH, S. (2006). Unsupervised learning with random forest predictors. *J. Comput. Graph. Statist.* **15** 118–138. MR2252461 https://doi.org/10.1198/106186006X94072

SPOHN, M.-L., NÄF, J., MICHEL, L. and MEINSHAUSEN, N. (2021). PKLM: A flexible MCAR test using classification. ArXiv preprint. Available at arXiv:2109.10150.

SRIPERUMBUDUR, B. K., FUKUMIZU, K., GRETTON, A., SCHÖLKOPF, B. and LANCKRIET, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Stat.* **6** 1550–1599. MR2988458 https://doi.org/10.1214/12-EJS722

STEKHOVEN, D. J. (2013). missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

STEKHOVEN, D. J. and BÜHLMANN, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28** 112–118.

THORARINSDOTTIR, T. L., GNEITING, T. and GISSIBL, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA J. Uncertain. Quantificat.* **1** 522–534. MR3283896 https://doi.org/10.1137/130907550

VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*, 2nd ed. CRC Press/CRC Press, Boca Raton, FL.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

WALJEE, A. K., MUKHERJEE, A., SINGAL, A. G., ZHANG, Y., WARREN, J., BALIS, U., MARRERO, J., ZHU, J. and HIGGINS, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* **3** e002847.

XU, X., XIA, L., ZHANG, Q., WU, S., WU, M. and LIU, H. (2020). The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Med. Res. Methodol.* **20** 1–16.

YOON, J., JORDON, J. and VAN DER SCHAAR, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **80** 5689–5698. PMLR.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2473–2493
https://doi.org/10.1214/23-AOAS1728

AN EFFICIENT DOUBLY-ROBUST IMPUTATION FRAMEWORK FOR LONGITUDINAL DROPOUT, WITH AN APPLICATION TO AN ALZHEIMER'S CLINICAL TRIAL

BY YUQI QIU[1,a] AND KAREN MESSER[2,b]

[1]*School of Statistics, East China Normal University,* [a]*yqqiu@fem.ecnu.edu.cn*
[2]*Division of Biostatistics and Bioinformatics, University of California San Diego,* [b]*kmesser@health.ucsd.edu*

We develop a novel doubly-robust (DR) imputation framework for longitudinal studies with monotone dropout, motivated by the informative dropout that is common in FDA-regulated trials for Alzheimer's disease. In this approach the missing data are first imputed using a doubly-robust augmented inverse probability weighting (AIPW) estimator; then the imputed completed data are substituted into a full-data estimating equation, and the estimate is obtained using standard software. The imputed completed data may be inspected and compared to the observed data, and standard model diagnostics are available. The same imputed completed data can be used for several different estimands, such as subgroup analyses in a clinical trial, allowing for reduced computation and increased consistency across analyses. We present two specific DR imputation estimators, AIPW-I and AIPW-S, study their theoretical properties, and investigate their performance by simulation. AIPW-S has substantially reduced computational burden, compared to many other DR estimators, at the cost of some loss of efficiency and the requirement of stronger assumptions. Simulation studies support the theoretical properties and good performance of the DR imputation framework. Importantly, we demonstrate their ability to address time-varying covariates, such as a time by treatment interaction. We illustrate using data from a large randomized Phase III trial, investigating the effect of donepezil in Alzheimer's disease, from the Alzheimer's Disease Cooperative Study (ADCS) group.

REFERENCES

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE (2010). Guideline on Missing Data in Confirmatory Clinical Trials. London: Eurpoean Medicines Agency.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

HSU, C.-H., HE, Y., LI, Y., LONG, Q. and FRIESE, R. (2016). Doubly robust multiple imputation using kernel-based techniques. *Biom. J.* **58** 588–606. MR3500562 https://doi.org/10.1002/bimj.201400256

INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE (2017). E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. Massachusetts Medical Society.

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

LONG, Q., HSU, C.-H. and LI, Y. (2012). Doubly robust nonparametric multiple imputation for ignorable missing data. *Statist. Sinica* **22** 149–172. MR2933171 https://doi.org/10.5705/ss.2010.069

PAIK, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *J. Amer. Statist. Assoc.* **92** 1320–1329.

PETERSEN, R. C., THOMAS, R. G., GRUNDMAN, M., BENNETT, D., DOODY, R., FERRIS, S., GALASKO, D., JIN, S., KAYE, J. et al. (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. *N. Engl. J. Med.* **352** 2379–2388.

Key words and phrases. Doubly-robust estimator, monotone dropouts, imputation methods, Alzheimer's disease, randomized trials.

QIU, Y. and MESSER, K. (2023). Supplement to "An efficient doubly-robust imputation framework for longitudinal dropout, with an application to an Alzheimer's clinical trial." https://doi.org/10.1214/23-AOAS1728SUPP

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

ROTNITZKY, A., LEI, Q., SUED, M. and ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456. MR2931264 https://doi.org/10.1093/biomet/ass013

SAS INSTITUTE INC. (2012). Online Documentation, SAS/STAT Version 9.3: Shared Concepts: LSMEANS Statement.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

SCHNITZER, M. E., LOK, J. J. and BOSCH, R. J. (2016). Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics* **17** 165–177. MR3449858 https://doi.org/10.1093/biostatistics/kxv028

SEAMAN, S. and COPAS, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Stat. Med.* **28** 937–955. MR2518358 https://doi.org/10.1002/sim.3520

SEAMAN, S. R. and VANSTEELANDT, S. (2018). Introduction to double robust methods for incomplete data. *Statist. Sci.* **33** 184–197. MR3797709 https://doi.org/10.1214/18-STS647

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

TSIATIS, A. A., DAVIDIAN, M. and CAO, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67** 536–545. MR2829022 https://doi.org/10.1111/j.1541-0420.2010.01476.x

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11. MR2306500 https://doi.org/10.2202/1557-4679.1043

XU, C., LI, Z., XUE, Y., ZHANG, L. and WANG, M. (2019). An R package for model fitting, model selection and the simulation for longitudinal data with dropout missingness. *Comm. Statist. Simulation Comput.* **48** 2812–2829. MR4001237 https://doi.org/10.1080/03610918.2018.1468457

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2494–2514
https://doi.org/10.1214/23-AOAS1729

A BAYESIAN GROWTH MIXTURE MODEL FOR COMPLEX SURVEY DATA: CLUSTERING POSTDISASTER PTSD TRAJECTORIES

BY REBECCA ANTHOPOLOS[1,a], QIXUAN CHEN[2,b], JOSEPH SEDRANSK[3,c], MARY THOMPSON[4,d], GANG MENG[5,e] AND SANDRO GALEA[6,f]

[1]*Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine,* [a]*rebecca.anthopolos@nyulangone.org*
[2]*Department of Biostatistics, Columbia University Mailman School of Public Health,* [b]*qc2138@cumc.columbia.edu*
[3]*Joint Program in Survey Methodology, University of Maryland,* [c]*jxs123@cwru.edu*
[4]*Department of Statistics and Actuarial Science, University of Waterloo,* [d]*methompson@uwaterloo.ca*
[5]*Department of Psychology, University of Waterloo,* [e]*gmeng@uwaterloo.ca*
[6]*School of Public Health, Boston University,* [f]*sgalea@bu.edu*

Research on growth mixture models (GMMs) for analyzing data from a complex sample survey is sparse. Existing methods use pseudo-likelihood in which survey weights are incorporated into the likelihood function, with variance estimated via linearization or resampling techniques. Despite popularity of the pseudo-likelihood approach, weighted estimation introduces the risk of efficiency loss. In this paper we propose a Bayesian GMM for complex survey data in which sample design features, such as stratification, clustering, and unequal probability of selection, are incorporated as covariates or hierarchical variance components. The Bayesian GMM can yield a reduction in bias in the estimation of regression coefficients when design features are associated with survey outcomes, and can lead to more efficient estimates than the pseudo-likelihood estimators when the design is noninformative. We develop an efficient Gibbs sampler that includes only closed-form full conditional distributions for model fitting. We present the results of a careful analysis of data from the Galveston Bay Recovery Study (GBRS) which used a stratified multi-stage cluster sample design. Using our proposed Bayesian GMM, we characterize longitudinal trajectories of post-traumatic stress disorder (PTSD) among residents of southeastern Texas in the aftermath of Hurricane Ike. We identify four clinically meaningful PTSD trajectory subgroups and characterize risk factors associated with subgroup membership. In the absence of existing software that can be used to implement our proposed Bayesian GMM for complex survey data, we built the `R` package `Bsvygmm` for model fitting, selection, and checking which can be downloaded from https://github.com/anthopolos/Bsvygmm.

REFERENCES

ABDALLA, A. and MICHAEL, S. (2019). Finite mixture of regression models for a stratified sample. *J. Stat. Comput. Simul.* **89** 2782–2800. MR3979730 https://doi.org/10.1080/00949655.2019.1636990

AITCHISON, J. and BENNETT, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika* **57** 253–262.

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

ANTHOPOLOS, R., CHEN, Q., SEDRANSK, J., THOMPSON, M., MENG, G. and GALEA, S. (2023). Supplement to "A Bayesian growth mixture model for complex survey data: Clustering postdisaster PTSD trajectories." https://doi.org/10.1214/23-AOAS1729SUPPA, https://doi.org/10.1214/23-AOAS1729SUPPB

ARIYO, O., LESAFFRE, E., VERBEKE, G., HUISMAN, M., HEYMANS, M. and TWISK, J. (2022). Bayesian model selection for multilevel mediation models. *Stat. Neerl.* **76** 219–235. MR4423266 https://doi.org/10.1111/stan.12256

Key words and phrases. Complex survey sample, Gibbs sampling, growth mixture model, post-traumatic stress disorder, spatial modeling.

ASPAROUHOV, T. (2005). Sampling weights in latent variable modeling. *Struct. Equ. Model.* **12** 411–434. MR2145982 https://doi.org/10.1207/s15328007sem1203_4

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

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746. MR1380811 https://doi.org/10.1093/biomet/82.4.733

BEST, N. G., SPIEGELHALTER, D. J., THOMAS, A. and BRAYNE, C. E. G. (1996). Bayesian analysis of realistically complex models. *J. Roy. Statist. Soc. Ser. A* **159** 323–342.

BONANNO, G. A. and DIMINICH, E. D. (2013). Annual research review: Positive adjustment to adversity–trajectories of minimal-impact resilience and emergent resilience. *J. Child Psychol. Psychiatry* **54** 378–401. https://doi.org/10.1111/jcpp.12021

BUNCH, D. S. (1991). Estimability in the multinomial probit model. *Transp. Res., Part B, Methodol.* **25** 1–12. MR1093617 https://doi.org/10.1016/0191-2615(91)90009-8

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal.* **1** 651–673. MR2282197 https://doi.org/10.1214/06-BA122

CHAMBERS, R. L. (2003). Introduction to Part A [Approaches to inference]. In *Analysis of Survey Data* (*Southampton*, 1999). *Wiley Ser. Surv. Methodol.* 13–28. Wiley, Chichester. MR1978841 https://doi.org/10.1002/0470867205

CHEN, Q., ELLIOTT, M. R. and LITTLE, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Surv. Methodol.* **36** 23–34.

CHEN, M. H., SHAO, Q. M. and IBRAHIM, J. G. (2012). *Monte Carlo Methods in Bayesian Computation*. Springer, Berlin.

CHEN, Q., ELLIOTT, M. R., HAZIZA, D., YANG, Y., GHOSH, M., LITTLE, R. J. A., SEDRANSK, J. and THOMPSON, M. (2017). Approaches to improving survey-weighted estimates. *Statist. Sci.* **32** 227–248. MR3648957 https://doi.org/10.1214/17-STS609

CHIB, S. and CARLIN, B. P. (1999). On MCMC sampling in hierarchical longitudinal models. *Stat. Comput.* **9** 17–26.

COLE, V. T. and BAUER, D. J. (2016). A note on the use of mixture models for individual prediction. *Struct. Equ. Model.* **23** 615–631. MR3508463 https://doi.org/10.1080/10705511.2016.1168266

DAGANZO, C. (1979). *Multinomial Probit. Economic Theory, Econometrics, and Mathematical Economics: The Theory and Its Application to Demand Forecasting*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London-Toronto, Ont.. MR0567139

ELLIOTT, M. R., GALLO, J. J., HAVE, T. R. T., BOGNER, H. R. and KATZ, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6** 119–143. https://doi.org/10.1093/biostatistics/kxh022

FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2012). *Applied Longitudinal Analysis* **998**. Wiley, New York.

FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635 https://doi.org/10.1198/016214502760047131

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer, New York. MR2265601

FRÜHWIRTH-SCHNATTER, S., CELEUX, G. and ROBERT, C. P., eds. (2019). *Handbook of Mixture Analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3889980

FRÜHWIRTH-SCHNATTER, S. and PYNE, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11** 317–336. https://doi.org/10.1093/biostatistics/kxp062

FRÜHWIRTH-SCHNATTER, S., TÜCHLER, R. and OTTER, T. (2004). Bayesian analysis of the heterogeneity model. *J. Bus. Econom. Statist.* **22** 2–15. MR2028204 https://doi.org/10.1198/073500103288619331

FULLER, W. A. (2011). *Sampling Statistics*. Wiley, Hoboken, NJ.

GARRETT, E. S. and ZEGER, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067. MR1815583 https://doi.org/10.1111/j.0006-341X.2000.01055.x

GEISSER, S. and EDDY, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153–160. MR0529531

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

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

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

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Taylor & Francis, London.

GRUEBNER, O., LOWE, S. R., TRACY, M., CERDÁ, M., JOSHI, S., NORRIS, F. H. and GALEA, S. (2016). The geography of mental health and general wellness in Galveston Bay after Hurricane Ike: A spatial epidemiologic study with longitudinal data. *Disaster Med. Public Health Prep.* **10** 261–273.

HAUSMAN, J. A. and WISE, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous perferences. *Econometrica* **46** 403–426. MR0465090 https://doi.org/10.2307/1913909

HEERINGA, S., WEST, B. and BERGLUND, P. (2015). Regression with complex samples. In *Regression Analysis and Causal Inference* (H. Best and C. Wolf, eds.) 225–248 11. Sage Publications, Thousand Oaks, CA.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). *Bayesian Survival Analysis. Springer Series in Statistics*. Springer, New York. MR1876598 https://doi.org/10.1007/978-1-4757-3447-8

KORN, E. L. and GRAUBARD, B. I. (1999). Sampling weights and imputation. In *Analysis of Health Surveys* 159–191. Wiley, New York.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LEÓN-NOVELO, L. G. and SAVITSKY, T. D. (2019). Fully Bayesian estimation under informative sampling. *Electron. J. Stat.* **13** 1608–1645. MR3939589 https://doi.org/10.1214/19-ejs1538

LIN, H., TURNBULL, B. W., MCCULLOCH, C. E. and SLATE, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Amer. Statist. Assoc.* **97** 53–65. MR1947272 https://doi.org/10.1198/016214502753479220

LITTLE, R. J. A. (1991). Inference with survey weights. *J. Off. Stat.* **7** 405–424.

LITTLE, R. J. (2003). The Bayesian approach to sample survey inference. In *Analysis of Survey Data* (*Southampton*, 1999). *Wiley Ser. Surv. Methodol.* 49–57. Wiley, Chichester. MR1978843 https://doi.org/10.1002/0470867205.ch4

LITTLE, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Amer. Statist. Assoc.* **99** 546–556. MR2109316 https://doi.org/10.1198/016214504000000467

LITTLE, R. J. A. and ZHENG, H. (2007). The Bayesian approach to the analysis of finite population surveys. In *Bayesian Statistics* 8. *Oxford Sci. Publ.* 283–302. Oxford Univ. Press, Oxford. MR2433197

LOHR, S. L. (2021). *Sampling: Design and Analysis*. CRC Press/CRC, Boca Raton, FL.

LOWE, S. R., JOSHI, S., PIETRZAK, R. H., GALEA, S. and CERDÁ, M. (2015). Mental health and general wellness in the aftermath of Hurricane Ike. *Soc. Sci. Med.* **124** 162–170.

MADDALA, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics. Econometric Society Monographs in Quantitative Economics* **3**. Cambridge Univ. Press, Cambridge. MR0799154 https://doi.org/10.1017/CBO9780511810176

MCCULLOCH, R. and ROSSI, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *J. Econometrics* **64** 207–240. MR1310524 https://doi.org/10.1016/0304-4076(94)90064-7

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

MUTHÉN, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychol. Methods* **8** 369–377.

MUTHÉN, L. K. and MUTHÉN, B. O. (2017). *Mplus User's Guide*, 8th ed. Muthén & Muthén, Los Angeles.

MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469. https://doi.org/10.1111/j.0006-341x.1999.00463.x

MUTHÉN, B., BROWN, C. H., MASYN, K., JO, B., KHOO, S. T., YANG, C. C., WANG, C. P., KELLAM, S. G., CARLIN, J. B. et al. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics* **3** 459–475.

NAGIN, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychol. Methods* **4** 139.

NASSERINEJAD, K., VAN ROSMALEN, J., DE KORT, W. and LESAFFRE, E. (2017). Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PLoS ONE* **12** e0168838. https://doi.org/10.1371/journal.pone.0168838

NEELON, B., SWAMY, G. K., BURGETTE, L. F. and MIRANDA, M. L. (2011). A Bayesian growth mixture model to examine maternal hypertension and birth outcomes. *Stat. Med.* **30** 2721–2735. MR2843175 https://doi.org/10.1002/sim.4291

NING, L. and LUO, W. (2018). Class identification efficacy in piecewise GMM with unknown turning points. *J. Exp. Educ.* **86** 282–307.

NORRIS, F. H., TRACY, M. and GALEA, S. (2009). Looking for resilience: Understanding the longitudinal trajectories of responses to stress. *Soc. Sci. Med.* **68** 2190–2198.

NORRIS, F. H., FRIEDMAN, M. J., WATSON, P. J., BYRNE, C. M. and KANIASTY, K. (2002). 60,000 disaster victims speak: Part I. An empirical review of the empirical literature, 1981–2001. *Psychiatry* **65** 207–239.

NATIONAL INSTITUTE OF MENTAL HEALTH (2016). Post-Traumatic Stress Disorder. Mental Health Information, Health Topics.

PAPASTAMOULIS, P. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw.* **69**.

PATTERSON, B. H., DAYTON, C. M. and GRAUBARD, B. I. (2002). Latent class analysis of complex sample survey data: Application to dietary data. *J. Amer. Statist. Assoc.* **97** 721–741. MR1941406 https://doi.org/10.1198/016214502388618465

PFEFFERMANN, D. (1996). The use of sampling weights for survey data analysis. *Stat. Methods Med. Res.* **5** 239–261.

PFEFFERMANN, D. and RAO, C. R., eds. (2009) *Sample Surveys*: *Inference and Analysis*. *Handbook of Statistics* **29**. Elsevier/North-Holland, Amsterdam.

PROUST-LIMA, C., SÉNE, M., TAYLOR, J. M. G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.* **23** 74–90. MR3190688 https://doi.org/10.1177/0962280212445839

RABE-HESKETH, S. and SKRONDAL, A. (2006). Multilevel modelling of complex survey data. *J. Roy. Statist. Soc. Ser. A* **169** 805–827. MR2291345 https://doi.org/10.1111/j.1467-985X.2006.00426.x

RAO, J. N. K. and WU, C.-F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241. MR0941020

RICE, H. (2016). Hurricane Ike Worst Storm in Decades. Houston Chronicle.

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

SI, Y., PALTA, M. and SMITH, M. (2020). Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records. *Ann. Appl. Stat.* **14** 1903–1924. MR4194253 https://doi.org/10.1214/20-AOAS1378

SI, Y., PILLAI, N. S. and GELMAN, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Anal.* **10** 605–625. MR3420817 https://doi.org/10.1214/14-BA924

SKINNER, C. J. (2003). Introduction to Part B [Categorical response data]. In *Analysis of Survey Data* (*Southampton*, 1999). *Wiley Ser. Surv. Methodol.* 75–84. Wiley, Chichester. MR1978845 https://doi.org/10.1002/0470867205

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

STATACORP (2019). mprobit – multinomial probit regression. In *Stata* 16 *Base Reference Manual* 1626–1632. Stata Press, College Station, TX.

STEELE, R. J. and RAFTERY, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. *Frontiers of Statistical Decision Making and Bayesian Analysis* **2** 113–130.

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

SUNG, C., HAALAND, B., HWANG, Y. and SIYUAN, L. (2019). A clustered Gaussian process model for computer experiments. Preprint. Available at arXiv:1911.04602.

TRAIN, K. E. (2009). *Discrete Choice Methods with Simulation*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2519514 https://doi.org/10.1017/CBO9780511805271

VALLIANT, R., ADAMS, T. and WAGNER, J. (2009). Sample Design Documentation Galveston Bay Recovery Survey 2008-2009 Technical Report Survey Research Operations, Production Sampling Group, Univ. Michigan Survey Research Center Ann Arbor, MI.

VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91** 217–221.

VERMUNT, J. K. and MAGIDSON, J. (2007). Latent class analysis with sampling weights: A maximum-likelihood approach. *Sociol. Methods Res.* **36** 87–111. MR2393664 https://doi.org/10.1177/0049124107301965

WANG, Z., KIM, J. K. and YANG, S. (2018). Approximate Bayesian inference under informative sampling. *Biometrika* **105** 91–102. MR3768867 https://doi.org/10.1093/biomet/asx073

WEDEL, M., HOFSTEDE, F. and STEENKAMP, J. E. M. (1998). Mixture model analysis of complex samples. *J. Classification* **15** 225–44.

ZHAO, L., FENG, D., NEELON, B. and BUYSE, M. (2015). Evaluation of treatment efficacy using a Bayesian mixture piecewise linear model of longitudinal biomarkers. *Stat. Med.* **34** 1733–1746. MR3334688 https://doi.org/10.1002/sim.6445

ZHENG, H. and LITTLE, R. J. A. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *J. Off. Stat.* **19** 99–117.

ZHOU, H., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *J. Off. Stat.* **32** 231–256. https://doi.org/10.1515/JOS-2016-0011

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2515–2532
https://doi.org/10.1214/23-AOAS1730

ESTIMATING HIV EPIDEMICS FOR SUBNATIONAL AREAS

BY LE BAO[1,a], XIAOYUE NIU[1,b], MARY MAHY[2,c] AND PETER D. GHYS[2,d]

[1]*Department of Statistics, Penn State University,* [a]*lebao@psu.edu,* [b]*xiaoyue@psu.edu*
[2]*Strategic Information and Evaluation Department, UNAIDS,* [c]*mahym@unaids.org,* [d]*ghysp@unaids.org*

As the global HIV pandemic enters its fifth decade, increasing numbers of countries use routine HIV testing among pregnant women to monitor their epidemics, allowing governments to look into the epidemics at a finer scale, for example, at subnational levels. Currently, the epidemic model that describes the dynamics of the spread of HIV consists of a set of differential equations and is applied independently to each subnational area. However, the availability of the data varies widely which leads to biased and unreliable estimates for areas with very few data points. We propose to overcome this issue by introducing dependence in the parameters across areas. The proposed method better reconstructs the epidemic trajectories than the independent model as shown in multiple countries in Sub-Saharan Africa. We also offer an approximate method for parameter estimation that is much less computationally burdensome than direct parameter estimation. Compared to direct parameter estimation from the dependent model, the approximate method provides competitive parameter estimation in simulations and the application of HIV subepidemic estimation.

REFERENCES

ALKEMA, L., RAFTERY, A. E. and CLARK, S. J. (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *Ann. Appl. Stat.* **1** 229–248. MR2393849 https://doi.org/10.1214/07-AOAS111

BAO, L. (2012). A new infectious disease model for estimating and projecting HIV/AIDS epidemics. *Sex. Transm. Infect.* **88** i58–i64.

BAO, L., NIU, X., MAHY, M. and GHYS, P. D (2023). Supplement to "Estimating HIV epidemics for subnational areas." https://doi.org/10.1214/23-AOAS1730SUPPA, https://doi.org/10.1214/23-AOAS1730SUPPB

BROWN, T., BAO, L., RAFTERY, A. E., SALOMON, J. A., BAGGALEY, R. F., STOVER, J. and GERLAND, P. (2010). EPP 2009: Bringing the UNAIDS estimation and projection package into the ART era. *Sex. Transm. Infect.* **86** ii3–ii10.

BROWN, T., BAO, L., EATON, J. W., HOGAN, D. R., MAHY, M., MARSH, K., MATHERS, B. M. and PUCKETT, R. (2014). Improvements in prevalence trend fitting and incidence estimation in EPP 2013. *AIDS* **28** S415–S425. https://doi.org/10.1097/QAD.0000000000000454

CALLEJA, J. M. G., JACOBSON, J., GARG, R., THUY, N., STENGAARD, A., ALONSO, M., ZIADY, H., MUKENGE, L., NTABANGANA, S. et al. (2010). Has the quality of serosurveillance in low-and middle-income countries improved since the last HIV estimates round in 2007? Status and trends through 2009. *Sex. Transm. Infect.* **86** ii35–ii42.

CASE, K. K., JOHNSON, L. F., MAHY, M., MARSH, K., SUPERVIE, V. and EATON, J. W. (2019). Summarizing the results and methods of the 2019 Joint United Nations Programme on HIV/AIDS HIV estimates. *AIDS* **33** S197.

CENTERS FOR DISEASE CONTROL (CDC) (1981). Pneumocystis pneumonia–Los Angeles. *Morb. Mort. Wkly. Rep.* **30** 250–252.

EATON, J. W., BROWN, T., PUCKETT, R., GLAUBIUS, R., MUTAI, K., BAO, L., SALOMON, J. A., STOVER, J., MAHY, M. et al. (2019). The estimation and projection package age-sex model and the R-hybrid model: New tools for estimating HIV incidence trends in sub-Saharan Africa. *AIDS* **33** S235.

LYERLA, R., GOUWS, E. and GARCIA-CALLEJA, J. M. (2008). The quality of sero-surveillance in low- and middle-income countries: Status and trends through 2007. *Sex. Transm. Infect.* **84** i85–i91. https://doi.org/10.1136/sti.2008.030593

MAHY, M., NZIMA, M., OGUNGBEMI, M. K., OGBANG, D. A., MORKA, M. C. and STOVER, J. (2014). Redefining the HIV epidemic in Nigeria: From national to state level. *AIDS* **28** S461–S468.

Key words and phrases. HIV epidemics, correlated model, dynamic system, importance sampling.

MARSH, K., MAHY, M., SALOMON, J. A. and HOGAN, D. R. (2014). Assessing and adjusting for differences between HIV prevalence estimates derived from national population-based surveys and antenatal care surveillance, with applications for spectrum 2013. *AIDS* **28** S497–S505. https://doi.org/10.1097/QAD.0000000000000453

RAFTERY, A. E. and BAO, L. (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics* **66** 1162–1173. MR2758504 https://doi.org/10.1111/j.1541-0420.2010.01399.x

STOVER, J., GLAUBIUS, R., MOFENSON, L., DUGDALE, C. M., DAVIES, M.-A., PATTEN, G. and YIANNOUTSOS, C. (2019). Updates to the spectrum/AIM model for estimating key HIV indicators at national and subnational levels. *AIDS* **33** S227.

UNAIDS (2013). *Location*, *Location*: *Connecting People Faster to HIV Services*. UNAIDS, Geneva, Switzerland.

UNAIDS (2014). *The GAP Report*. UNAIDS, Geneva, Switzerland.

UNAIDS (2020). *Seizing the Moment*: *Global AIDS Update Report*. UNAIDS, Geneva, Switzerland.

UNAIDS and WHO (2015). *Guidelines on Monitoring the Impact of the HIV Epidemic Using Population-Based Surveys*. UNAIDS, Geneva, Switzerland.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2533–2554
https://doi.org/10.1214/23-AOAS1731

JOINT MODELING OF PLAYING TIME AND PURCHASE PROPENSITY IN MASSIVELY MULTIPLAYER ONLINE ROLE-PLAYING GAMES USING CROSSED RANDOM EFFECTS

BY TRAMBAK BANERJEE[1,a], PENG LIU[2,b], GOURAB MUKHERJEE[3,c], SHANTANU DUTTA[4,d] AND HAI CHE[5,e]

[1]*Analytics, Information and Operations Management, University of Kansas,* [a]*trambak@ku.edu*
[2]*Department of Marketing, Santa Clara University,* [b]*pliu2@scu.edu*
[3]*Department of Data Sciences and Operations, University of Southern California,* [c]*gmukherj@marshall.usc.edu*
[4]*Department of Marketing, University of Southern California,* [d]*shantanu@marshall.usc.edu*
[5]*Department of Marketing, University of California, Riverside,* [e]*chehai@ucr.edu*

Massively multiplayer online role-playing games (MMORPGs) offer a unique blend of a personalized gaming experience and a platform for forging social connections. Managers of these digital products rely on predictions of key player responses, such as playing time and purchase propensity, to design timely interventions for promoting, engaging and monetizing their playing base. However, the longitudinal data associated with these MMORPGs not only exhibit a large set of potential predictors to choose from but often present several other distinctive characteristics that pose significant challenges in developing flexible statistical algorithms that can generate efficient predictions of future player activities. For instance, the existence of virtual communities or "guilds" in these games complicate prediction since players who are part of the same guild have correlated behaviors and the guilds themselves evolve over time and thus have a dynamic effect on the future playing behavior of its members. In this paper we develop a *crossed random effects joint modeling* (CREJM) framework for analyzing correlated player responses in MMORPGs. Contrary to existing methods that assume player independence, CREJM is flexible enough to incorporate both player dependence as well as time-varying guild effects on the future playing behavior of the guild members. On a large-scale data from a popular MMORPG, CREJM conducts simultaneous selection of fixed and random effects in high-dimensional penalized multivariate mixed models. We study the asymptotic properties of the variable selection procedure in CREJM and establish its selection consistency. Besides providing superior predictions of daily playing time and purchase propensity over competing methods, CREJM also predicts player correlations within each guild which are valuable for optimizing future promotional and reward policies for these virtual communities.

REFERENCES

BANERJEE, T., MUKHERJEE, G., DUTTA, S. and GHOSH, P. (2020). A large-scale constrained joint modeling approach for predicting user activity, engagement, and churn with application to freemium mobile games. *J. Amer. Statist. Assoc.* **115** 538–554. MR4107656 https://doi.org/10.1080/01621459.2019.1611584

BANERJEE, T., LIU, P., MUKHERJEE, G., DUTTA, S. and CHE, H. (2023a). Supplement to "Joint modeling of playing time and purchase propensity in massively multiplayer online role-playing games using crossed random effects." https://doi.org/10.1214/23-AOAS1731SUPPA

BANERJEE, T., LIU, P., MUKHERJEE, G., DUTTA, S. and CHE, H. (2023b). Source code for "Joint modeling of playing time and purchase propensity in massively multiplayer online role-playing games using crossed random effects." https://doi.org/10.1214/23-AOAS1731SUPPB

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325 https://doi.org/10.1093/biomet/asr054

Key words and phrases. Large-scale longitudinal data analysis, massively multiplayer online role-playing games, monetization of digital products, online communities, guilds, cross-classified random effect models.

BONDELL, H. D., KRISHNA, A. and GHOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66** 1069–1077. MR2758494 https://doi.org/10.1111/j.1541-0420.2010.01391.x

BORBORA, Z., SRIVASTAVA, J., HSU, K.-W. and WILLIAMS, D. (2011). Churn prediction in mmorpgs using player motivation theories and an ensemble approach. In 2011 *IEEE Third International Conference on Privacy, Security, Risk and Trust and* 2011 *IEEE Third International Conference on Social Computing* 157–164. IEEE, New York.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CAFRI, G. and FAN, J. (2018). Between-within effects in survival models with cross-classified clustering: Application to the evaluation of the effectiveness of medical devices. *Stat. Methods Med. Res.* **27** 312–319. MR3745672 https://doi.org/10.1177/0962280216628561

CAFRI, G., HEDEKER, D. and AARONS, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychol. Methods* **20** 407.

CANDÈS, E. J., WAKIN, M. B. and BOYD, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611 https://doi.org/10.1007/s00041-008-9045-x

CHEN, J. and CHEN, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statist. Sinica* **22** 555–574. MR2954352 https://doi.org/10.5705/ss.2010.216

CISION (2020). Implications of COVID-19 on the global role playing games market. News: September 2020. Available at https://www.prnewswire.com/news-releases/implications-of-covid-19-on-the-global-role-playing-games-market-301139710.html.

CLEMENTS, R. (2012). RPGs took over every video game genre. Available at https://www.ign.com/articles/2012/12/12/rpgs-took-over-every-video-game-genre.

DFCINTELLIGENCE (2020). Global video game consumer segmentation. Available at https://www.dfcint.com/product/video-game-consumer-segmentation-2/.

FAN, Y. and LI, R. (2012). Variable selection in linear mixed effects models. *Ann. Statist.* **40** 2043–2068. MR3059076 https://doi.org/10.1214/12-AOS1028

GAO, K. (2017). Scalable estimation and inference for massive linear mixed models with crossed random effects. Ph.D. thesis, Stanford University. MR4257222

GAO, K. and OWEN, A. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electron. J. Stat.* **11** 1235–1296. MR3635913 https://doi.org/10.1214/17-EJS1236

GAO, K. and OWEN, A. B. (2020). Estimation and inference for very large linear mixed effects models. *Statist. Sinica* **30** 1741–1771. MR4260743 https://doi.org/10.5705/ss.202018.0029

GHOSH, S., HASTIE, T. and OWEN, A. B. (2022). Backfitting for large scale crossed random effects regressions. *Ann. Statist.* **50** 560–583. MR4382028 https://doi.org/10.1214/21-aos2121

HACKMAN, J. R. and VIDMAR, N. (1970). Effects of size and task type on group performance and member reactions. *Sociom.* 37–54.

HUANG, Y., JASIN, S. and MANCHANDA, P. (2019). "Level up": Leveraging skill and engagement to maximize player game-play in online video games. *Inf. Syst. Res.* **30** 927–947.

HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statist. Sinica* **27** 501–518. MR3674683

HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2017b). Joint selection in mixed models using regularized PQL. *J. Amer. Statist. Assoc.* **112** 1323–1333. MR3735380 https://doi.org/10.1080/01621459.2016.1215989

HUI, F. K. C., MÜLLER, S. and WELSH, A. H. (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *J. Amer. Statist. Assoc.* **113** 1759–1769. MR3902244 https://doi.org/10.1080/01621459.2017.1371026

IBRAHIM, J. G., ZHU, H., GARCIA, R. I. and GUO, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67** 495–503. MR2829018 https://doi.org/10.1111/j.1541-0420.2010.01463.x

JIANG, J., RAO, J. S., GU, Z. and NGUYEN, T. (2008). Fence methods for mixed model selection. *Ann. Statist.* **36** 1669–1692. MR2435452 https://doi.org/10.1214/07-AOS517

JIN, W. and SUN, Y. (2015). Understanding the antecedents of virtual product purchase in MMORPG: An integrative perspective of social presence and user engagement. In *PACIS* 191.

KANG, J., KO, I. and KO, Y. (2009). The impact of social support of guild members and psychological factors on flow and game loyalty in MMORPG. In 2009 42*nd Hawaii International Conference on System Sciences* 1–9. IEEE, New York.

KHANNA, R., ZHANG, L., AGARWAL, D. and CHEN, B.-C. (2013). Parallel matrix factorization for binary response. In 2013 *IEEE International Conference on Big Data* 430–438. IEEE, New York.

KOREN, Y., BELL, R. and VOLINSKY, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30–37.

KUMAR, V. (2014). Making "freemium" work. *Harv. Bus. Rev.* **92** 27–29.
LE, C. M. and LI, T. (2022). Linear regression and its inference on noisy network-linked data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1851–1885. MR4515560
LI, T., LEVINA, E. and ZHU, J. (2019). Prediction models for network-linked data. *Ann. Appl. Stat.* **13** 132–164. MR3937424 https://doi.org/10.1214/18-AOAS1205
LIN, B., PANG, Z. and JIANG, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *J. Comput. Graph. Statist.* **22** 341–355. MR3173718 https://doi.org/10.1080/10618600.2012.681219
LU, C., LIN, Z. and YAN, S. (2015). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Trans. Image Process.* **24** 646–654. MR3301260 https://doi.org/10.1109/TIP.2014.2380155
MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105 https://doi.org/10.2307/2291460
PAN, J. and HUANG, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Stat. Comput.* **24** 725–738. MR3229693 https://doi.org/10.1007/s11222-013-9398-0
PAPASPILIOPOULOS, O., ROBERTS, G. O. and ZANELLA, G. (2020). Scalable inference for crossed random effects models. *Biometrika* **107** 25–40. MR4064138 https://doi.org/10.1093/biomet/asz058
PARK, E., RISHIKA, R., JANAKIRAMAN, R., HOUSTON, M. B. and YOO, B. (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *J. Mark.* **82** 93–114.
PENG, H. and LU, Y. (2012). Model selection in linear mixed effect models. *J. Multivariate Anal.* **109** 109–129. MR2922858 https://doi.org/10.1016/j.jmva.2012.02.005
RABE-HESKETH, S., SKRONDAL, A., PICKLES, A. et al. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J.* **2** 1–21.
RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *J. Educ. Stat.* **18** 321–349.
RAUDENBUSH, S. W. and BRYK, A. S. (2002). *Hierarchical Linear Models*: *Applications and Data Analysis Methods* 1. Sage, Thousand Oaks.
RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256 https://doi.org/10.1111/j.1541-0420.2010.01546.x
RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*: *With Applications in R*. CRC Press, Boca Raton.
SCHELLDORFER, J., MEIER, L. and BÜHLMANN, P. (2014). GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *J. Comput. Graph. Statist.* **23** 460–477. MR3215820 https://doi.org/10.1080/10618600.2013.773239
TERLUTTER, R. and CAPELLA, M. L. (2013). The gamification of advertising: Analysis and research directions of in-game advertising, advergames, and advertising in social network games. *J. Advert.* **42** 95–112.
TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567
WEI, G. C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.
WEI, Y., ZHANG, W., YANG, S. and CHEN, X. (2019). Online communities and social network structure. Available at SSRN 3420525.
ZHANG, C., PHANG, C. W., WU, Q. and LUO, X. (2017). Nonlinear effects of social connections and interactions on individual goal attainment and spending: Evidences from online gaming markets. *J. Mark.* **81** 132–155.
ZHAO, Y.-B. and KOČVARA, M. (2015). A new computational method for the sparsest solutions to systems of linear equations. *SIAM J. Optim.* **25** 1110–1134. MR3357641 https://doi.org/10.1137/140968240

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2555–2573
https://doi.org/10.1214/23-AOAS1732

STRUCTURE LEARNING FOR ZERO-INFLATED COUNTS WITH AN APPLICATION TO SINGLE-CELL RNA SEQUENCING DATA

BY THI KIM HUE NGUYEN[1,a], KOEN VAN DEN BERGE[2,c], MONICA CHIOGNA[3,d] AND DAVIDE RISSO[1,b]

[1]*Department of Statistical Sciences, Universiy of Padova,* [a]*nguyen@stat.unipd.it,* [b]*davide.risso@unipd.it*
[2]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University,* [c]*koenvdberge@berkeley.edu*
[3]*Department of Statistical Sciences, University of Bologna,* [d]*monica.chiogna2@unibo.it*

The problem of estimating the structure of a graph from observed data is of growing interest in the context of high-throughput genomic data and single-cell RNA sequencing in particular. These, however, are challenging applications, since the data consist of high-dimensional counts with high variance and overabundance of zeros. Here we present a general framework for learning the structure of a graph from single-cell RNA-seq data, based on the zero-inflated negative binomial distribution. We demonstrate with simulations that our approach is able to retrieve the structure of a graph in a variety of settings, and we show the utility of the approach on real data.

REFERENCES

ABEGAZ, F. and WIT, E. (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Stat. Neerl.* **69** 419–441. MR3414705 https://doi.org/10.1111/stan.12066

ALLEN, G. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.

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

BRANN, D. H., TSUKAHARA, T., WEINREB, C., LIPOVSEK, M., VAN DEN BERGE, K., GONG, B., CHANCE, R., MACAULAY, I. C., CHOU, H. J. et al. (2020). Non-neuronal expression of Sars-CoV-2 entry genes in the olfactory system suggests mechanisms underlying Covid-19-associated anosmia. *Sci. Adv.* **6** eabc5801.

BRYAN, A. H. (2020). HiveR: 2D and 3D Hive Plots for R. R package version 0.3.63.

BULT, C. J., BLAKE, J. A., SMITH, C. L., KADIN, J. A., RICHARDSON, J. E. and THE MOUSE GENOME DATABASE GROUP (2019). Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47** D801–D806.

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

CHEN, D., YAN, W., FU, L. Y. and KAUFMANN, K. (2018). Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat. Commun.* **9** 4534. https://doi.org/10.1038/s41467-018-06772-3

COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

CSARDI, G., NEPUSZ, T. et al. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems* **1695** 1–9.

CSISZÁR, I. and TALATA, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145. MR2275237 https://doi.org/10.1214/009053605000000912

DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393. https://doi.org/10.1146/annurev-statistics-060116-053803

FLETCHER, R. B., DAS, D., GADYE, L., STREET, K. N., BAUDHUIN, A., WAGNER, A., COLE, M. B., FLORES, Q., CHOI, Y. G. et al. (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell Stem Cell* **20** 817–830.

GADYE, L., DAS, D., SANCHEZ, M. A., STREET, K., BAUDHUIN, A., WAGNER, A., COLE, M. B., CHOI, Y. G., YOSEF, N. et al. (2017). Injury activates transient olfactory stem cell states with diverse lineage capacities. *Cell Stem Cell* **21** 775–790.

Key words and phrases. Structure learning, graphical models, zero-inflated counts, single-cell RNA-seq.

GALLOPIN, M., RAU, A. and JAFFRÉZIC, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE* **8** e77503.

GONZÁLEZ, B., DENZEL, S., MACK, B., CONRAD, M. and GIRES, O. (2009). EpCAM is involved in maintenance of the murine embryonic stem cell phenotype. *Stem Cells* **27** 1782–1791. https://doi.org/10.1002/stem.97

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264. https://doi.org/10.1093/biostatistics/4.2.249

ISLAM, S., ZEISEL, A., JOOST, S., MANNO, G. L., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166. https://doi.org/10.1038/nmeth.2772

JO, A., DENDULURI, S., ZHANG, B., WANG, Z., YIN, L., YAN, Z., KANG, R., SHI, L. L., MOK, J. et al. (2014). The versatile functions of Sox9 in development, stem cells, and human diseases. *Genes Dis.* **1** 149–161.

JUNBAI, W., LEO, W. K. C. and JAN, D. (2005). New probabilistic graphical models for genetic regulatory networks studies. *J. Biomed. Inform.* **38** 443–455. https://doi.org/10.1016/j.jbi.2005.04.003

KOLODZIEJCZYK, A. A., KIM, J. K., SVENSSON, V., MARIONI, J. C. and TEICHMANN, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58** 610–620.

KRZYWINSKI, M., BIROL, I., JONES, S. J. and MARRA, M. A. (2012). Hive plots—rational approach to visualizing networks. *Brief. Bioinform.* **13** 627–644.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York. MR1419991

LIBERZON, A., BIRGER, C., THORVALDSDÓTTIR, H., GHANDI, M., MESIROV, J. P. and TAMAYO, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* **1** 417–425. https://doi.org/10.1016/j.cels.2015.12.004

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440.

LIU, K., LIN, B., ZHAO, M., YANG, X., CHEN, M., GAO, A., LIU, F., QUE, J. and LAN, X. (2013). The multiple roles for Sox2 in stem cell maintenance and tumorigenesis. *Cell. Signal.* **25** 1264–1271.

MARIN NAVARRO, A., PRONK, R. J., VAN DER GEEST, A. T., OLIYNYK, G., NORDGREN, A., ARSENIAN-HENRIKSSON, M., FALK, A. and WILHELM, M. (2020). P53 controls genomic stability and temporal differentiation of human neural stem cells and affects neural organization in human brain organoids. *Cell Death & Disease* **11** 52. https://doi.org/10.1038/s41419-019-2208-7

MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *Ann. Appl. Stat.* **13** 848–873. MR3963555 https://doi.org/10.1214/18-AOAS1213

MEYERS, E. A. and KESSLER, J. A. (2017). TGF-β family signaling in neural and neuronal differentiation, development, and function. *Cold Spring Harb. Perspect. Biol.* **9** a022244. https://doi.org/10.1101/cshperspect.a022244

NGUYEN, T. K. H. and CHIOGNA, M. (2021). Structure learning of undirected graphical models for count data. *J. Mach. Learn. Res.* **22** Paper No. 50, 53. MR4253743

NGUYEN, T. K, VAN DEN BERGE, K., CHIOGNA, M. and RISSO, D. (2023). Supplement to "Structure learning for zero-inflated counts with an application to single-cell RNA sequencing data." https://doi.org/10.1214/23-AOAS1732SUPPA, https://doi.org/10.1214/23-AOAS1732SUPPB, https://doi.org/10.1214/23-AOAS1732SUPPC

PEÑA, J. M. (2008). Learning Gaussian graphical models of gene networks with false discovery rate control. In *Evolutionary Computation*, *Machine Learning and Data Mining in Bioinformatics*. (J. H. Moore and E. Marchiori, ed.). *Lecture Notes in Computer Science* **4973**. Springer, Berlin.

RIEGE, K., KRETZMER, H., SAHM, A., MCDADE, S. S., HOFFMANN, S. and FISCHER, M. (2020). Dissecting the DNA binding landscape and gene regulatory network of p63 and p53. *eLife* **9**. https://doi.org/10.7554/eLife.63266

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 1–17.

SARKAR, A. and STEPHENS, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53** 770–777. https://doi.org/10.1038/s41588-021-00873-4

SCHÄFER, J. and STRIMMER, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21** 754–764. https://doi.org/10.1093/bioinformatics/bti062

SENOO, M., PINTO, F., CRUM, C. P. and MCKEON, F. (2007). p63 is essential for the proliferative potential of stem cells in stratified epithelia. *Cell* **129** 523–536. https://doi.org/10.1016/j.cell.2007.02.045

SIKDAR, S. and DATTA, S. (2017). A novel statistical approach for identification of the master regulator transcription factor. *BMC Bioinform.* **18** 79. https://doi.org/10.1186/s12859-017-1499-x

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550. https://doi.org/10.1073/pnas.0506580102

SVENSSON, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38** 1–4.

TANG, C., WANG, M., WANG, P., WANG, L., WU, Q. and GUO, W. (2019). Neural stem cells behave as a functional niche for the maturation of newborn neurons through the secretion of PTN. *Neuron* **101** 32–44.

TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20** 1–16.

TRAAG, V. A., WALTMAN, L. and VAN ECK, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9** 5233. https://doi.org/10.1038/s41598-019-41695-z

VAN DE WIEL, M. A., LEDAY, G. G. R., PARDO, L., RUE, H., VAN DER VAART, A. W. and VAN WIERINGEN, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14** 113–128. https://doi.org/10.1093/biostatistics/kxs031

VIETH, B., ZIEGENHAIN, C., PAREKH, S., ENARD, W. and HELLMANN, I. (2017). powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33** 3486–3488. https://doi.org/10.1093/bioinformatics/btx435

WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. https://doi.org/10.1038/nrg2484

WANG, S. and ROHE, K. (2016). Discussion of "Coauthorship and citation networks for statisticians" [MR3592033]. *Ann. Appl. Stat.* **10** 1820–1826. MR3592035 https://doi.org/10.1214/16-AOAS977

WANG, W., LIU, W., WANG, Y., ZHOU, L., TANG, X. and LUO, H. (2011). Notch signaling regulates neuroepithelial stem cell maintenance and neuroblast formation in Drosophila optic lobe development. *Dev. Psychobiol.* **350** 414–428. https://doi.org/10.1016/J.YDBIO.2010.12.002

YANG, Z. and HO, Y.-Y. (2022). Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics* **78** 766–776. MR4450593 https://doi.org/10.1111/biom.13457

YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129 https://doi.org/10.1214/11-AOAS494

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2574–2595
https://doi.org/10.1214/23-AOAS1733

BAYESIAN INFERENCE AND DYNAMIC PREDICTION FOR MULTIVARIATE LONGITUDINAL AND SURVIVAL DATA

BY HAOTIAN ZOU[1,a], DONGLIN ZENG[1,b], LUO XIAO[2,c] AND SHENG LUO[3,d]

[1]*Department of Biostatistics, University of North Carolina at Chapel Hill,* [a]*haotian@live.unc.edu,* [b]*dzeng@email.unc.edu*
[2]*Department of Statistics, North Carolina State University,* [c]*lxiao5@ncsu.edu*
[3]*Department of Biostatistics and Bioinformatics, Duke University,* [d]*sheng.luo@duke.edu*

Alzheimer's disease (AD) is a complex neurological disorder impairing multiple domains such as cognition and daily functions. To better understand the disease and its progression, many AD research studies collect multiple longitudinal outcomes that are strongly predictive of the onset of AD dementia. We propose a joint model based on a multivariate functional mixed model framework (referred to as MFMM-JM) that simultaneously models the multiple longitudinal outcomes and the time to dementia onset. We develop six functional forms to fully investigate the complex association between longitudinal outcomes and dementia onset. Moreover, we use the Bayesian methods for statistical inference and develop a dynamic prediction framework that provides accurate personalized predictions of disease progressions based on new subject-specific data. We apply the proposed MFMM-JM to two large ongoing AD studies, the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC), and identify the functional forms with the best predictive performance. Our method is also validated by extensive simulation studies with five settings.

REFERENCES

ANDRINOPOULOU, E.-R. and RIZOPOULOS, D. (2016). Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. *Stat. Med.* **35** 4813–4823. MR3554995 https://doi.org/10.1002/sim.7027

BALADANDAYUTHAPANI, V., JI, Y., TALLURI, R., NIETO-BARAJAS, L. E. and MORRIS, J. S. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *J. Amer. Statist. Assoc.* **105** 1358–1375. Supplementary materials available online. MR2796556 https://doi.org/10.1198/jasa.2010.ap09250

BALLARD, C., GAUTHIER, S., CORBETT, A., BRAYNE, C., AARSLAND, D. and JONES, E. (2011). Alzheimer's disease. *Lancet* **377** 1019–1031.

BESSER, L., KUKULL, W., KNOPMAN, D. S., CHUI, H., GALASKO, D., WEINTRAUB, S., JICHA, G., CARLSSON, C., BURNS, J. et al. (2018). Version 3 of the national Alzheimer's coordinating center's uniform data set. *Alzheimer Dis. Assoc. Disord.*

BREIJYEH, Z. and KARAMAN, R. (2020). Comprehensive review on Alzheimer's disease: Causes and treatment. *Molecules* **25**. https://doi.org/10.3390/molecules25245789

BROWN, E. R. (2009). Assessing the association between trends in a biomarker and risk of event with an application in pediatric HIV/AIDS. *Ann. Appl. Stat.* **3** 1163–1182. MR2750391 https://doi.org/10.1214/09-AOAS251

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

CORDER, E. H., SAUNDERS, A. M., STRITTMATTER, W. J., SCHMECHEL, D. E., GASKELL, P. C., SMALL, G. W., ROSES, A. D., HAINES, J. L. and PERICAK-VANCE, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261** 921–923. https://doi.org/10.1126/science.8346443

DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S. and PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.* **3** 458–488. MR2668715 https://doi.org/10.1214/08-AOAS206

Key words and phrases. Alzheimer's disease, multivariate longitudinal data, functional mixed model, joint model, Bayesian method, dynamic prediction.

FENG, S., WOLFE, R. A. and PORT, F. K. (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using Poisson variance structures. *J. Amer. Statist. Assoc.* **100** 728–735. MR2206989 https://doi.org/10.1198/016214505000000123

GALIMARD, J.-E., CHEVRET, S., PROTOPOPESCU, C. and RESCHE-RIGON, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat. Med.* **35** 2907–2920. MR3528233 https://doi.org/10.1002/sim.6902

GOLDSMITH, J. and KITAGO, T. (2016). Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 215–236. MR3456686 https://doi.org/10.1111/rssc.12115

GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18** 2529–2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x

HAMMON, A. and ZINN, S. (2020). Multiple imputation of binary multilevel missing not at random data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 547–564. MR4098961

HICKEY, G. L., PHILIPSON, P., JORGENSEN, A. and KOLAMUNNAGE-DONA, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Med. Res. Methodol.* **16** 1–15.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

KANG, J., REICH, B. J. and STAICU, A.-M. (2018). Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105** 165–184. MR3768872 https://doi.org/10.1093/biomet/asx075

LAWLESS, J. and ZHAN, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canad. J. Statist.* **26** 549–565.

LI, L., GREENE, T. and HU, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat. Methods Med. Res.* **27** 2264–2278. MR3825906 https://doi.org/10.1177/0962280216680239

LI, K. and LUO, S. (2019). Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease. *Stat. Methods Med. Res.* **28** 327–342. MR3903744 https://doi.org/10.1177/0962280217722177

LI, C., XIAO, L. and LUO, S. (2020). Fast covariance estimation for multivariate sparse functional data. *Stat* **9** e245, 18. MR4116315

LI, C., XIAO, L. and LUO, S. (2022). Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease. *Biometrics* **78** 435–447. MR4450566 https://doi.org/10.1111/biom.13427

LI, M., NG, T. P., KUA, E. H. and KO, S. M. (2006). Brief informant screening test for mild cognitive impairment and early Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* **21** 392–402. https://doi.org/10.1159/000092808

LI, K., CHAN, W., DOODY, R. S., QUINN, J., LUO, S. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2017). Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J. Alzheimer's Dis.* **58** 361–371. https://doi.org/10.3233/JAD-161201

LONG, J. D. and MILLS, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Med. Res. Methodol.* **18** 1–15.

MAUFF, K., STEYERBERG, E., KARDYS, I., BOERSMA, E. and RIZOPOULOS, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. *Stat. Comput.* **30** 999–1014. MR4108688 https://doi.org/10.1007/s11222-020-09927-9

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.

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., VANNUCCI, M., BROWN, P. J. and CARROLL, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Amer. Statist. Assoc.* **98** 573–597. With comments and a rejoinder by the authors. MR2011673 https://doi.org/10.1198/016214503000000422

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu. Rev. Stat. Appl.* **6** 223–240. MR3939519 https://doi.org/10.1146/annurev-statistics-030718-105048

PETERSEN, R. C. (2016). Mild cognitive impairment. *Continuum* (*Minneap Minn*) **22** 404–418. https://doi.org/10.1212/CON.0000000000000313

RAJAN, K. B., WEUVE, J., BARNES, L. L., MCANINCH, E. A., WILSON, R. S. and EVANS, D. A. (2021). Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020–2060). *Alzheimer's Dement*.

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256 https://doi.org/10.1111/j.1541-0420.2010.01546.x

RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*: *With Applications in R*. CRC Press.

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959 https://doi.org/10.1002/sim.4205

RIZOPOULOS, D., HATFIELD, L. A., CARLIN, B. P. and TAKKENBERG, J. J. M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *J. Amer. Statist. Assoc.* **109** 1385–1397. MR3293598 https://doi.org/10.1080/01621459.2014.931236

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

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16** 667–718. Includes comments and discussions by seven discussants and a rejoinder by the authors. MR4298989 https://doi.org/10.1214/20-ba1221

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WANG, J. and LUO, S. (2017). Bayesian multivariate augmented Beta rectangular regression models for patient-reported outcomes and survival data. *Stat. Methods Med. Res.* **26** 1684–1699. MR3687172 https://doi.org/10.1177/0962280215586010

WANG, G., LIU, S., HAN, F. and DI, C. (2021). Robust functional principal component analysis via functional pairwise spatial signs. arXiv preprint arXiv:2101.06415.

WEINER, M. W., VEITCH, D. P., AISEN, P. S., BECKETT, L. A., CAIRNS, N. J., CEDARBAUM, J., GREEN, R. C., HARVEY, D., JACK, C. R. et al. (2015). 2014 update of the Alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's Dement.* **11** e1–e120.

WENG, S.-C., CHANG, Y.-C. and CHEN, C.-M. (2022). Joint analysis of longitudinal and interval-censored failure time data. *Comm. Statist. Simulation Comput.* **51** 5333–5349. MR4491685 https://doi.org/10.1080/03610918.2020.1770284

YAO, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statist. Sinica* **17** 965–983. MR2408647

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561 https://doi.org/10.1198/016214504000001745

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

ZOU, H., ZENG, D., XIAO, L. and LUO, S. (2023a). Supplementary material for "Bayesian inference and dynamic prediction for multivariate longitudinal and survival data." https://doi.org/10.1214/23-AOAS1733SUPPA

ZOU, H., ZENG, D., XIAO, L. and LUO, S. (2023b). Code for "Bayesian inference and dynamic prediction for multivariate longitudinal and survival data." https://doi.org/10.1214/23-AOAS1733SUPPB

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2596–2618
https://doi.org/10.1214/23-AOAS1734

ESTIMATING GARCH(1, 1) IN THE PRESENCE OF MISSING DATA

BY DAMIEN C. H. WEE[a], FENG CHEN[b] AND WILLIAM T. M. DUNSMUIR[c]

School of Mathematics and Statistics, UNSW Sydney, [a]*dchwee@gmail.com,* [b]*feng.chen@unsw.edu.au,* [c]*w.dunsmuir@unsw.edu.au*

Maximum likelihood estimation of the famous GARCH(1, 1) model is generally straightforward, given the full observation series. However, when some observations are missing, the marginal likelihood of the observed data is intractable in most cases of interest, also intractable is the likelihood from temporally aggregated data. For both these problems, we propose to approximate the intractable likelihoods through sequential Monte Carlo (SMC). The SMC approximation is done in a smooth manner so that the resulting approximate likelihoods can be numerically optimized to obtain parameter estimates. In the case with data aggregation, the use of SMC is made possible by a novel state space representation of the aggregated GARCH series. Through extensive simulation experiments, the proposed method is found to be computationally feasible and produce more accurate estimators of the model parameters compared with other recently published methods, especially in the case with aggregated data. In addition, the Hessian matrix of the minus logarithm of the approximate likelihood can be inverted to produce fairly accurate standard error estimates. The proposed methodology is applied to the analysis of time series data on several exchange-traded funds on the Australian Stock Exchange with missing prices, due to interruptions such as scheduled trading holidays.

REFERENCES

BLASQUES, F., GORGI, P. and KOOPMAN, S. J. (2021). Missing observations in observation-driven time series models. *J. Econometrics* **221** 542–568. MR4215038 https://doi.org/10.1016/j.jeconom.2020.07.043

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051 https://doi.org/10.1016/0304-4076(86)90063-1

BONDON, P. and BAHAMONDE, N. (2012). Least squares estimation of ARCH models with missing observations. *J. Time Series Anal.* **33** 880–891. MR2991906 https://doi.org/10.1111/j.1467-9892.2012.00803.x

CASCONE, M. H. and HOTTA, L. K. (2019). Quasi-maximum likelihood estimation of GARCH models in the presence of missing values. *J. Stat. Comput. Simul.* **89** 292–314. MR3882041 https://doi.org/10.1080/00949655.2018.1546860

CREAL, D., SCHWAAB, B. and KOOPMAN, S. J. (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Rev. Econ. Stat.* **96** 898–915.

CRIPPS, E. and DUNSMUIR, W. (2003). Modeling the variability of Sydney harbor wind measurements. *J. Appl. Meteorol.* **42** 1131–1138.

DEL MORAL, P. (1996). Nonlinear filtering: Interacting particle solution. *Markov Process. Related Fields* **2** 555–579. MR1431187

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

EDDELBUETTEL, D. and FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

FRANCQ, C. and THIEU, L. Q. (2019). QML inference for volatility models with covariates. *Econometric Theory* **35** 37–72. MR3904171 https://doi.org/10.1017/S0266466617000512

GÁBOR, D. K., SÁVAI, M. and UDVARI, B. (2017). Missing data bias on a selective hedging strategy. *J. Compet.* **9** 15–19.

GLOSTEN, L., JAGANNATHAN, R. and RUNKLE, D. (1992). On the relation between the expected value and the volatility of nominal excess return on stocks. *J. Finance* **46** 1779–1801.

Key words and phrases. Hidden Markov model, incomplete data, intractable likelihood, sequential Monte Carlo, volatility modelling.

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE Proc. F* **140** 107–113.

HAN, H. and KRISTENSEN, D. (2014). Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates. *J. Bus. Econom. Statist.* **32** 416–429. MR3238595 https://doi.org/10.1080/07350015.2014.897954

HANSEN, P. R. and LUNDE, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)? *J. Appl. Econometrics* **20** 873–889. MR2223415 https://doi.org/10.1002/jae.800

HIGGINS, P. (2014). GDPNow: A model for GDP "nowcasting". Federal Reserve Bank of Atlanta: Working Paper Series.

KANTAS, N., DOUCET, A., SINGH, S. and MACIEJOWSKI, J. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. *IFAC Proc. Vol.* **42** 774–785.

KELLY, M. and CLARK, S. (2011). Returns in trading versus non-trading hours: The difference is day and night. *J. Asset Manag.* **12** 132–145.

LOU, D., POLK, C. and SKOURAS, S. (2019). A tug of war: Overnight versus intraday expected returns. *J. Financ. Econ.* **134** 192–213.

MALIK, S. and PITT, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *J. Econometrics* **165** 190–209. MR2846644 https://doi.org/10.1016/j.jeconom.2011.07.006

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

NELSON, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59** 347–370. MR1097532 https://doi.org/10.2307/2938260

OSSANDÓN, S. and BAHAMONDE, N. (2013). A new nonlinear formulation for GARCH models. *C. R. Math. Acad. Sci. Paris* **351** 235–239. MR3089685 https://doi.org/10.1016/j.crma.2013.02.014

PITT, M. K., MALIK, S. and DOUCET, A. (2014). Simulated likelihood inference for stochastic volatility models using continuous particle filtering. *Ann. Inst. Statist. Math.* **66** 527–552. MR3211873 https://doi.org/10.1007/s10463-014-0456-y

ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525 https://doi.org/10.1214/aoms/1177729394

SUCARRAT, G. and ESCRIBANO, A. (2018). Estimation of log-GARCH models in the presence of zero returns. *Eur. J. Finance* **24** 809–827.

SUCARRAT, G. and GRØNNEBERG, S. (2022). Risk estimation with a time-varying probability of zero returns. *J. Financ. Econom.* **20** 278–309.

R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

TERASVIRTA, T. (2009). Introduction to univariate GARCH models. In *Handbook of Financial Time Series* 17–42. Springer, Berlin.

WEE, D. C., CHEN, F. and DUNSMUIR, W. T. (2023). Supplement to "Estimating GARCH(1, 1) in the presence of missing data." https://doi.org/10.1214/23-AOAS1734SUPP

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2619–2638
https://doi.org/10.1214/23-AOAS1735

SNIP: AN ADAPTATION OF SORTED NEIGHBORHOOD METHODS FOR DEDUPLICATING PEDIGREE DATA

BY THEODORE HUANG[a], MATTHEW PLOENZKE[b] AND DANIELLE BRAUN[c]

Department of Biostatistics, Harvard T.H. Chan School of Public Health, [a]*thuang@ds.dfci.harvard.edu,*
[b]*matthew.ploenzke@gmail.com,* [c]*dbraun@mail.harvard.edu*

Pedigree data contain family history information that is used to analyze hereditary diseases. These clinical data sets may contain duplicate records due to the same family visiting a clinic multiple times or a clinician entering multiple versions of the family for testing purposes. Inferences drawn from the data or using them for training or validation without removing the duplicates could lead to invalid conclusions, and hence identifying the duplicates is essential. Since family structures can be complex, direct application of existing deduplication algorithms may not be straightforward. We first motivate the importance of deduplication by examining the impact of pedigree duplicates on model performance when training and validating a familial risk prediction model. We then introduce an unsupervised algorithm, which we call SNIP (**S**orted **N**e**I**ghborhood for **P**edigrees), that builds on the sorted neighborhood method to find efficiently and to classify pair comparisons by leveraging the inherent hierarchical nature of the pedigrees. We conduct a simulation study to assess the performance of the algorithm and find parameter configurations where the algorithm is able to accurately detect the duplicates. We then apply the method to data from the Risk Service, which includes over 300,000 pedigrees at high risk of hereditary cancers, and uncover large clusters of potential duplicate families. After removing 104,520 pedigrees (33% of original data), the resulting Risk Service data set can now be used for future analysis, training, and validation. The algorithm is available as an R package snipR at https://github.com/bayesmendel/snipR.

REFERENCES

ANTON-CULVER, H., ZIOGAS, A., BOWEN, D., FINKELSTEIN, D., GRIFFIN, C., HANSON, J., ISAACS, C., KASTEN-SPORTES, C., MINEAU, G. et al. (2003). The cancer genetics network: Recruitment results and pilot studies. *Publ. Health Genom.* **6** 171–177.

BELIN, T. R. and RUBIN, D. B. (1995). A method for calibrating false-match rates in record linkage. *J. Amer. Statist. Assoc.* **90** 694–707.

BILENKO, M., KAMATH, B. and MOONEY, R. J. (2006). Adaptive blocking: Learning to scale up record linkage. In *Data Mining*, 2006. *ICDM'*06. *Sixth International Conference on* 87–96. IEEE, New York.

CALADO, P., HERSCHEL, M. and LEITÃO, L. (2010). An overview of XML duplicate detection algorithms. *Soft Computing in XML Data Management* 193–224.

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** Art. 21. MR2101490 https://doi.org/10.2202/1544-6115.1063

CHIPMAN, J., DROHAN, B., BLACKFORD, A., PARMIGIANI, G., HUGHES, K. and BOSINOFF, P. (2013). Providing access to risk prediction tools via the HL7 XML-formatted risk web service. *Breast Cancer Res. Treat.* **140** 187–193.

CSARDI, G., NEPUSZ, T. et al. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems* **1695** 1–9.

DHIVYABHARATHI, G. and KUMARESAN, S. (2016). A survey on duplicate record detection in real world data. In *Advanced Computing and Communication Systems* (*ICACCS*), 2016 3*rd International Conference on* **1** 1–5. IEEE, New York.

Key words and phrases. Deduplication, pedigree data, entity matching, sorted neighborhood.

DRAISBACH, U., NAUMANN, F., SZOTT, S. and WONNEBERG, O. (2012). Adaptive windows for duplicate detection. In *Data Engineering* (*ICDE*), 2012 *IEEE* 28*th International Conference on* 1073–1083. IEEE, New York.

ELMAGARMID, A. K., IPEIROTIS, P. G. and VERYKIOS, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.* **19** 1–16.

FREEDMAN, A. N., SLATTERY, M. L., BALLARD-BARBASH, R., WILLIS, G., CANN, B. J., PEE, D., GAIL, M. H. and PFEIFFER, R. M. (2009). Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J. Clin. Oncol.* **27** 686–693. https://doi.org/10.1200/JCO.2008.17.4797

HERNÁNDEZ, M. A. and STOLFO, S. J. (1995). The merge/purge problem for large databases. *ACM Sigmod Record* **24** 127–138.

HERZOG, J. S., CHAVARRI-GUERRA, Y., CASTILLO, D., ABUGATTAS, J., VILLARREAL-GARZA, C., SAND, S., CLAGUE-DEHART, J., ALVAREZ-GÓMEZ, R. M., WEGMAN-OSTROSKY, T. et al. (2021). Genetic epidemiology of BRCA1-and BRCA2-associated cancer across Latin America. *npj Breast Cancer* **7** 1–8.

HUANG, J., ERTEKIN, S. and GILES, C. L. (2006). Efficient name disambiguation for large-scale databases. In *European Conference on Principles of Data Mining and Knowledge Discovery* 536–544. Springer, Berlin.

HUANG, T., PLOENZKE, M. and BRAUN, D. (2023). Supplement to "SNIP: An adaptation of sorted neighborhood methods for deduplicating pedigree data." https://doi.org/10.1214/23-AOAS1735SUPPA, https://doi.org/10.1214/23-AOAS1735SUPPB

IDOS, G. E., KURIAN, A. W., RICKER, C., STURGEON, D., CULVER, J. O., KINGHAM, K. E., KOFF, R., CHUN, N. M., ROWE-TEETER, C. et al. (2019). Multicenter prospective cohort study of the diagnostic yield and patient experience of multiplex gene panel testing for hereditary cancer risk. *JCO Precision Oncology* **3** 1–12.

IVIE, S., PIXTON, B. and GIRAUD-CARRIER, C. (2007). Metric-based data mining model for genealogical record linkage. In 2007 *IEEE International Conference on Information Reuse and Integration* 538–543. IEEE, New York.

KOLB, L., THOR, A. and RAHM, E. (2012). Multi-pass sorted neighborhood blocking with MapReduce. *Computer Science-Research and Development* **27** 45–63.

KÖPCKE, H. and RAHM, E. (2010). Frameworks for entity matching: A comparison. *Data Knowl. Eng.* **69** 197–210.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. MR1973781 https://doi.org/10.1198/016214501750332956

LEE, G., LIANG, J. W., ZHANG, Q., HUANG, T., CHOIRAT, C., PARMIGANI, G. and BRAUN, D. (2021). Multi-syndrome, multi-gene risk modeling for individuals with a family history of cancer with the novel R package PanelPRO. *eLife* **10** e68699.

MANNING, C. D., RAGHAVAN, P. and SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge.

MENESTRINA, D., WHANG, S. E. and GARCIA-MOLINA, H. (2010). Evaluating entity resolution results. *Proc. VLDB Endow.* **3** 208–219.

PAPADAKIS, G., KOUTRIKA, G., PALPANAS, T. and NEJDL, W. (2014). Meta-blocking: Taking entity resolutionto the next level. *IEEE Trans. Knowl. Data Eng.* **26** 1946–1960.

PIXTON, B. and GIRAUD-CARRIER, C. (2005). MAL4: 6-using data mining for record linkage. In *Proceedings of the* 5*th Annual Workshop on Technology for Family History and Genealogical Research* Citeseer.

PIXTON, B. and GIRAUD-CARRIER, C. (2006). Using structured neural networks for record linkage. In *Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research*.

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.

TYRER, J., DUFFY, S. W. and CUZICK, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23** 1111–1130. https://doi.org/10.1002/sim.1668

WALDRON, L., RIESTER, M., RAMOS, M., PARMIGIANI, G. and BIRRER, M. (2016). The Doppelgänger effect: Hidden duplicates in databases of transcriptome profiles. *J. Natl. Cancer Inst.* **108**. https://doi.org/10.1093/jnci/djw146

WILLIAMS-BLANGERO, S. and BLANGERO, J. (2006). Collection of pedigree data for genetic analysis in isolate populations. *Hum. Biol.* **78** 89–101. https://doi.org/10.1353/hub.2006.0023

YAN, S., LEE, D., KAN, M.-Y. and GILES, L. C. (2007). Adaptive sorted neighborhood methods for efficient record linkage. In *Proceedings of the* 7*th ACM/IEEE-CS Joint Conference on Digital Libraries* 185–194. ACM, New York.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2639–2658
https://doi.org/10.1214/23-AOAS1736

A HORSESHOE MIXTURE MODEL FOR BAYESIAN SCREENING WITH AN APPLICATION TO LIGHT SHEET FLUORESCENCE MICROSCOPY IN BRAIN IMAGING

BY FRANCESCO DENTI[1,a], RICARDO AZEVEDO[2,b], CHELSIE LO[2,c], DAMIAN G. WHEELER[3,e], SUNIL P. GANDHI[2,d], MICHELE GUINDANI[4,f] AND BABAK SHAHBABA[5,g]

[1]*Department of Statistics, Università Cattolica del Sacro Cuore,* [a]*francesco.denti@unicatt.it*
[2]*Department of Neurobiology and Behavior, University of California, Irvine,* [b]*azevedor@uci.edu,* [c]*lochelsie@gmail.com,* [d]*sunil.gandhi@uci.edu*
[3]*Translucence Biosystems, Inc,* [e]*damian@translucencebio.com*
[4]*Department of Biostatistcs, University of California, Los Angeles,* [f]*mguindani@g.ucla.edu*
[5]*Department of Statistics, University of California, Irvine,* [g]*babaks@uci.edu*

In this paper we focus on identifying differentially activated brain regions using a light sheet fluorescence microscopy—a recently developed technique for whole-brain imaging. Most existing statistical methods solve this problem by partitioning the brain regions into two classes: significantly and nonsignificantly activated. However, for the brain imaging problem at the center of our study, such binary grouping may provide overly simplistic discoveries by filtering out weak but important signals that are typically adulterated by the noise present in the data. To overcome this limitation, we introduce a new Bayesian approach that allows classifying the brain regions into several tiers with varying degrees of *relevance*. Our approach is based on a combination of shrinkage priors, widely used in regression and multiple hypothesis testing problems, and mixture models, commonly used in model-based clustering. In contrast to the existing regularizing prior distributions, which use either the spike-and-slab prior or continuous scale mixtures, our class of priors is based on a *discrete mixture of continuous scale mixtures* and devises a cluster shrinkage version of the horseshoe prior. As a result, our approach provides a more general setting for Bayesian sparse estimation, drastically reduces the number of shrinkage parameters needed, and creates a framework for sharing information across units of interest. We show that this approach leads to more biologically meaningful and interpretable results in our brain imaging problem, since it allows the discrimination between active and inactive regions, while at the same time ranking the discoveries into clusters representing tiers of similar importance.

REFERENCES

ANDERMANN, M. L., KERLIN, A. M., ROUMIS, D. K., GLICKFELD, L. L. and REID, R. C. (2011). Functional specialization of mouse higher visual cortical areas. *Neuron* **72** 1025–1039. https://doi.org/10.1016/j.neuron.2011.11.013

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

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2019). Lasso meets horseshoe: A survey. *Statist. Sci.* **34** 405–427. MR4017521 https://doi.org/10.1214/19-STS700

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. T. (2021). The horseshoe-like regularization for feature subset selection. *Sankhya B* **83** 185–214. MR4256316 https://doi.org/10.1007/s13571-019-00217-7

BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048 https://doi.org/10.1080/01621459.2014.960967

Key words and phrases. Bayesian inference, variable selection, mixture models, neuroscience.

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

DENTI, F., AZEVEDO, R., LO, C., WHEELER, D. G., GANDHI, S. P., GUINDANI, M. and SHAHBABA, B. (2023). Supplement to "A horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging." https://doi.org/10.1214/23-AOAS1736SUPPA, https://doi.org/10.1214/23-AOAS1736SUPPB

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089 https://doi.org/10.1214/009053606000001460

FINEGOLD, M. and DRTON, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. MR2840186 https://doi.org/10.1214/10-AOAS410

FINEGOLD, M. and DRTON, M. (2014). Robust Bayesian graphical modeling using Dirichlet t-distributions. *Bayesian Anal.* **9** 521–550. MR3256052 https://doi.org/10.1214/13-BA856

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

GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440 https://doi.org/10.1214/10-BA507

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

HRVATIN, S., HOCHBAUM, D. R., NAGY, M. A., CICCONET, M., ROBERTSON, K., CHEADLE, L., ZILIONIS, R., RATNER, A., BORGES-MONROY, R. et al. (2018). Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21** 120–129.

HÜBENER, M. (2003). Mouse visual cortex. *Curr. Opin. Neurobiol.* **13** 413–420. https://doi.org/10.1016/s0959-4388(03)00102-8

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158 https://doi.org/10.1214/009053604000001147

JOHNDROW, J., ORENSTEIN, P. and BHATTACHARYA, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *J. Mach. Learn. Res.* **21** 73. MR4095352

LIN, Y., BLOODGOOD, B. L., HAUSER, J. L., LAPAN, A. D., KOON, A. C., KIM, T. K., HU, L. S., MALIK, A. N. and GREENBERG, M. E. (2008). Activity-dependent regulation of inhibitory synapse development by Npas4. *Nature* **455** 1198–1204.

MAKALIC, E. and SCHMIDT, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. ArXiv Preprint 1–18.

MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. MR3439375 https://doi.org/10.1007/s11222-014-9500-2

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

MCSHANE, B. B., GAL, D., GELMAN, A., ROBERT, C. and TACKETT, J. L. (2019). Abandon statistical significance. *Amer. Statist.* **73** 235–245. MR3925729 https://doi.org/10.1080/00031305.2018.1527253

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE (2019). *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC.

MEDVEDOVIC, M., YEUNG, K. Y. and BUMGARNER, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20** 1222–1232. https://doi.org/10.1093/bioinformatics/bth068

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

NEVILLE, S. E., ORMEROD, J. T. and WAND, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. MR3263115 https://doi.org/10.1214/14-EJS910

NIELL, C. M. and STRYKER, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65** 472–479.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001 https://doi.org/10.1198/016214508000000337

PIIRONEN, J. and VEHTARI, A. (2017a). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. MR3738204 https://doi.org/10.1214/17-EJS1337SI

PIIRONEN, J. and VEHTARI, A. (2017b). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics*, *AISTATS* 2017 1–9.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2014). The Bayesian bridge. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 713–733. MR3248673 https://doi.org/10.1111/rssb.12042

POLSON, N. G., SCOTT, J. G., CLARKE, B. and SEVERINSKI, C. (2012). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9** 1–30.

RAMAMOORTHI, K., FROPF, R., BELFORT, G. M., FITZMAURICE, H. L., MCKINNEY, R. M., NEVE, R. L., OTTO, T. and LIN, Y. (2011). Npas4 regulates a transcriptional program in CA3 required for contextual memory formation. *Science* **334** 1669–1675.

RENIER, N., ADAMS, E. L., KIRST, C., WU, Z., AZEVEDO, R., KOHL, J., AUTRY, A. E., KADIRI, L., UMADEVI VENKATARAJU, K. et al. (2016). Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* **165** 1789–1802.

RICHARDSON, D. S. and LICHTMAN, J. W. (2015). Clarifying tissue clearing. *Cell* **162** 246–257. https://doi.org/10.1016/j.cell.2015.06.067

ROČKOVÁ, V. and GEORGE, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.* **113** 431–444. MR3803476 https://doi.org/10.1080/01621459.2016.1260469

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

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

SHAHBABA, B. and JOHNSON, W. O. (2013). Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes. *Stat. Med.* **32** 2114–2126. MR3067360 https://doi.org/10.1002/sim.5680

SHENG, M. and GREENBERG, M. E. (1990). The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* **4** 477–485. https://doi.org/10.1016/0896-6273(90)90106-p

SUN, X. and LIN, Y. (2016). Npas4: Linking neuronal activity to memory. *Trends Neurosci.* **39** 264–275.

SUNKIN, S. M., NG, L., LAU, C., DOLBEARE, T., GILBERT, T. L., THOMPSON, C. L., HAWRYLYCZ, M. and DANG, C. (2013). Allen brain atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41** 996–1008.

TADESSE, M. and VANNUCCI, M. (2021). *Handbook of Bayesian Variable Selection*. CRC Press/CRC, New York.

TUKEY, J. W. (1993). Tightening the clinical trial. *Control. Clin. Trials* **14** 266–285. https://doi.org/10.1016/0197-2456(93)90225-3

VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985 https://doi.org/10.1214/17-BA1065

WASSERSTEIN, R. L., SCHIRM, A. L. and LAZAR, N. A. (2019). Moving to a world beyond "$p < 0.05$". *Amer. Statist.* **73** 1–19.

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2659–2679
https://doi.org/10.1214/23-AOAS1739

USING PREDICTABILITY TO IMPROVE MATCHING OF URBAN LOCATIONS IN PHILADELPHIA

BY COLMAN HUMPHREY[a], RYAN GROSS[b], DYLAN S. SMALL[c] AND SHANE T. JENSEN[d]

Department of Statistics, The Wharton School, University of Pennsylvania , [a]*colmanhumphrey@gmail.com*, [b]*rzgross@wharton.upenn.edu*, [c]*dsmall@wharton.upenn.edu*, [d]*stjensen@wharton.upenn.edu*

Motivated by theories in urban planning and criminology, we use high-resolution data to investigate the relationship between crime and the built environment in the City of Philadelphia. We develop a novel and flexible matching framework that uses the predictability of the treatment variable within matched pairs to empirically inform both the differential weighting of covariates in the matching as well as the selection of the number of matched pairs to create. We use this matching framework for a series of comparisons, each involving matched pairs of Philadelphia intersections that are highly similar on a set of covariates but restricted to differ on a single aspect of the built environment. Our predictability-based matching framework includes data-driven decisions about differential weighting of covariates and the number of matched pairs to create, which is beneficial in our setting as our urban comparisons involve a large number of potential intersections and a large set of covariates to be balanced. In these comparisons we find substantial heterogeneity in the relationships between crime and different aspects of the built environment as well as some empirical support for historical theories.

REFERENCES

BRANAS, C. C., CHENEY, R. A., MACDONALD, J. M., TAM, V. W., JACKSON, T. D. and HAVE, T. R. T. (2011). A difference-in-differences analysis of health, safety, and greening vacant urban space. *Amer. J. Epidemiol*. **174** 1296–1306.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78** 1–3.

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist*. **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

BUJA, A., STUETZLE, W. and SHEN, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November 3.

CHEN, H. and SMALL, D. S. (2022). New multivariate tests for assessing covariate balance in matched observational studies. *Biometrics* **78** 202–213. MR4408581 https://doi.org/10.1111/biom.13395

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

COLEMAN, C. and MOYNIHAN, J. (1996). *Understanding Crime Data*: *Haunted by the Dark Figure*. Open Univ. Press, Maidenhead.

COZENS, P. M., SAVILLE, G. and HILLIER, D. (2005). Crime prevention through environmental design (CPTED): A review and modern bibliography. *Prop. Manag*. **23** 328–356.

CUI, J., JENSEN, S. T. and MACDONALD, J. (2022). The effects of vacant lot greening and the impact of land use and business presence on crime. *Environ. Plan. B*: *Urban Anal. City Sci*. **49** 1147–1158.

DESHPANDE, S. K., HASEGAWA, R. B., RABINOWITZ, A. R., WHYTE, J., ROAN, C. L., TABATABAEI, A., BAIOCCHI, M., KARLAWISH, J. H., MASTER, C. L. et al. (2017). Association of playing high school football with cognition and mental health later in life. *JAMA Neurol*. **74** 909–918. https://doi.org/10.1001/jamaneurol.2017.1317

GAGNON-BARTSCH, J. and SHEM-TOV, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *Ann. Appl. Stat*. **13** 1464–1483. MR4019146 https://doi.org/10.1214/19-AOAS1241

GU, X. and ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph. Statist*. **2** 405–420.

Key words and phrases. Matching, predictability, urban data, crime.

GURM, H. S., HOSMAN, C., SHARE, D., MOSCUCCI, M. and HANSEN, B. B. (2013). Comparative safety of vascular closure devices and manual closure among patients having percutaneous coronary intervention. *Ann. Intern. Med.* **159** 660–666.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *Amer. Statist.* **64** 299–309. MR2758561 https://doi.org/10.1198/tast.2010.09210

HUMPHREY, C., GROSS, R., SMALL, D. and JENSEN, S. T. (2023). Supplement to "Using predictability to improve matching of urban locations in Philadelphia." https://doi.org/10.1214/23-AOAS1739SUPPA, https://doi.org/10.1214/23-AOAS1739SUPPB

JACOBS, J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

JEFFERY, C. R. (1871). *Crime Prevention Through Environmental Design*. Sage Publications, Beverly Hills.

LAI, S. B. S., SHAHRI, N. H. N. B. M., MOHAMAD, M. B., RAHMAN, H. A. B. A. and RAMBLI, A. B. (2021). Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data. *Math. Stat.* **9** 379–385.

LEE, B. K., LESSLER, J. and STUART, E. A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* **29** 337–346. MR2750549 https://doi.org/10.1002/sim.3782

LOHR, S. L. (2019). *Measuring Crime*: *Behind the Statistics*. CRC Press, Boca Raton, FL.

LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. MR2899649 https://doi.org/10.1198/tast.2011.08294

MACDONALD, J. (2015). Community design and crime: The impact of housing and the built environment. *Crime Justice* **44** 333–383.

MACDONALD, J., BRANAS, C. and STOKES, R. (2019). *Changing Places*: *The Science and Art of New Urban Planning*. Princeton Univ. Press, Princeton.

MAHALANOBIS, P. C. (1936). On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **2** 49–55.

NEWMAN, O. (1972). *Defensible Space*: *Crime Prevention Through Urban Design*. MacMillian, New York.

PAPADIMITRIOU, C. H. and STEIGLITZ, K. (1998). *Combinatorial Optimization*: *Algorithms and Complexity*. Dover, Mineola, NY. MR1637890

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

ROSENBAUM, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist.* **21** 57–71. MR2913356 https://doi.org/10.1198/jcgs.2011.09219

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

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29** 159–184.

SEKHON, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R.

SILBER, J. H., ROSENBAUM, P. R., MCHUGH, M. D., LUDWIG, J. M., SMITH, H. L., NIKNAM, B. A., EVEN-SHOSHAN, O., FLEISHER, L. A., KELZ, R. R. et al. (2016). Comparison of the value of nursing work environments in hospitals across different levels of patient risk. *JAMA Surg.* **151** 527–536. https://doi.org/10.1001/jamasurg.2015.4908

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

TAYLOR, R. B. (1988). *Human Territorial Functioning*. Cambridge Univ. Press, Cambridge.

TU, C. (2019). Comparison of various machine learning algorithms for estimating generalized propensity score. *J. Stat. Comput. Simul.* **89** 708–719. MR3904596 https://doi.org/10.1080/00949655.2019.1571059

WESTREICH, D., LESSLER, J. and FUNK, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* **63** 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

WILCOX, P. and CULLEN, F. T. (2018). Situational opportunity theories of crime. *Annu. Rev. Criminol.* **1** 123–148.

WILSON, J. Q. and KELLING, G. L. (1982). Broken windows: The police and neighborhood safety. *Atl. Mon.* **29** 29–38.

YU, R. (2021). Evaluating and improving a matched comparison of antidepressants and bone density. *Biometrics* **77** 1276–1288. MR4357837 https://doi.org/10.1111/biom.13374

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2680–2699
https://doi.org/10.1214/23-AOAS1741

A SEMIPARAMETRIC PROMOTION TIME CURE MODEL WITH SUPPORT VECTOR MACHINE

BY SUVRA PAL[a] AND WISDOM ASELISEWINE[b]

Department of Mathematics, University of Texas at Arlington, [a]*suvra.pal@uta.edu,* [b]*wxa5616@mavs.uta.edu*

The promotion time cure rate model (PCM) is an extensively studied model for the analysis of time-to-event data in the presence of a cured subgroup. There are several strategies proposed in the literature to model the latency part of PCM. However, there aren't many strategies proposed to investigate the effects of covariates on the incidence part of PCM. In this regard most existing studies assume the boundary separating the cured and noncured subjects with respect to the covariates to be linear. As such, they can only capture simple effects of the covariates on the cured/noncured probability. In this manuscript we propose a new promotion time cure model that uses the support vector machine (SVM) to model the incidence part. The proposed model inherits the features of the SVM and provides flexibility in capturing nonlinearity in the data. To the best of our knowledge, this is the first work that integrates the SVM with PCM model. For the estimation of model parameters, we develop an expectation maximization algorithm where we make use of the sequential minimal optimization technique together with the Platt scaling method to obtain the posterior probabilities of cured/uncured. A detailed simulation study shows that the proposed model outperforms the existing logistic regression-based PCM model as well as the spline regression-based PCM model, which is also known to capture nonlinearity in the data. This is true in terms of bias and mean square error of different quantities of interest and also in terms of predictive and classification accuracies of cure. Finally, we illustrate the applicability and superiority of our model using the data from a study on leukemia patients who went through bone marrow transplantation.

REFERENCES

ASANO, J., HIRAKAWA, A. and HAMADA, C. (2014). Assessing the prediction accuracy of cure in the Cox proportional hazards cure model: An application to breast cancer data. *Pharm. Stat.* **13** 357–363.

BALAKRISHNAN, N., KOUTRAS, M. V., MILIENOS, F. S. and PAL, S. (2016). Piecewise linear approximations for cure rate models and associated inferential issues. *Methodol. Comput. Appl. Probab.* **18** 937–966. MR3564846 https://doi.org/10.1007/s11009-015-9477-0

BALAKRISHNAN, N. and PAL, S. (2012). EM algorithm-based likelihood estimation for some cure rate models. *J. Stat. Theory Pract.* **6** 698–724. MR3196574 https://doi.org/10.1080/15598608.2012.719803

BALAKRISHNAN, N. and PAL, S. (2013). Lognormal lifetimes and likelihood-based inference for flexible cure rate models based on COM-Poisson family. *Comput. Statist. Data Anal.* **67** 41–67. MR3079583 https://doi.org/10.1016/j.csda.2013.04.018

BALAKRISHNAN, N. and PAL, S. (2015a). Likelihood inference for flexible cure rate models with gamma lifetimes. *Comm. Statist. Theory Methods* **44** 4007–4048. MR3406329 https://doi.org/10.1080/03610926.2014.964807

BALAKRISHNAN, N. and PAL, S. (2015b). An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood-and information-based methods. *Comput. Statist.* **30** 151–189. MR3334716 https://doi.org/10.1007/s00180-014-0527-9

BALAKRISHNAN, N. and PAL, S. (2016). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes. *Stat. Methods Med. Res.* **25** 1535–1563. MR3541112 https://doi.org/10.1177/0962280213491641

BERKSON, J. and GAGE, R. P. (1952). Survival curve for cancer patients following treatment. *J. Amer. Statist. Assoc.* **47** 501–515.

Key words and phrases. Machine learning, multiple imputation, sequential minimal optimization, EM algorithm, ROC.

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. Ser. B* **11** 15–53.

BROWN, E. R. and IBRAHIM, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59** 686–693. MR2004274 https://doi.org/10.1111/1541-0420.00079

CAI, C., ZOU, Y., PENG, Y. and ZHANG, J. (2012). smcure: An R-package for estimating semi-parametric mixture cure models. *Comput. Methods Programs Biomed.* **108** 1255–1260.

CHEN, M.-H. and IBRAHIM, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57** 43–52. MR1833290 https://doi.org/10.1111/j.0006-341X.2001.00043.x

CHEN, M.-H., IBRAHIM, J. G. and SINHA, D. (1999). A new Bayesian model for survival data with a surviving fraction. *J. Amer. Statist. Assoc.* **94** 909–919. MR1723307 https://doi.org/10.2307/2670006

CHEN, T. and DU, P. (2018). Promotion time cure rate model with nonparametric form of covariate effects. *Stat. Med.* **37** 1625–1635. MR3787977 https://doi.org/10.1002/sim.7597

COONER, F., BANERJEE, S., CARLIN, B. P. and SINHA, D. (2007). Flexible cure rate modeling under latent activation schemes. *J. Amer. Statist. Assoc.* **102** 560–572. MR2370853 https://doi.org/10.1198/016214507000000112

COPELAN, E. A., BIGGS, J. C., THOMPSON, J. M., CRILLEY, P., SZER, J., KLEIN, J. P., KAPOOR, N., AVALOS, B. R., CUNNINGHAM, I. et al. (1991). Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2. *Blood* **78** 838–843.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

FLOYD, C. E., LO, J. Y., YUN, A. J., SULLIVAN, D. C. and KORNGUTH, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. *Cancer* **74** 2944–2948. https://doi.org/10.1002/1097-0142(19941201)74:11<2944::aid-cncr2820741109>3.0.co;2-f

FOUODO, C. J. K., KONIG, I. R., WEIHS, C., ZIEGLER, A. and WRIGHT, M. N. (2018). Support vector machines for survival analysis with R. *R J.* **10** 412–423.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. *Springer Series in Statistics*. Springer, New York. MR1851606 https://doi.org/10.1007/978-0-387-21606-5

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 https://doi.org/10.1214/08-AOAS169

KOKONENDJI, C. C., MIZÈRE, D. and BALAKRISHNAN, N. (2008). Connections of the Poisson weight function to overdispersion and underdispersion. *J. Statist. Plann. Inference* **138** 1287–1296. MR2388011 https://doi.org/10.1016/j.jspi.2007.05.028

LI, P., PENG, Y., JIANG, P. and DONG, Q. (2020). A support vector machine based semiparametric mixture cure model. *Comput. Statist.* **35** 931–945. MR4133104 https://doi.org/10.1007/s00180-019-00931-w

LIU, H. and SHEN, Y. (2009). A semiparametric regression cure model for interval-censored data. *J. Amer. Statist. Assoc.* **104** 1168–1178. MR2750242 https://doi.org/10.1198/jasa.2009.tm07494

MALLER, R. A. and ZHOU, X. (1996). *Survival Analysis with Long-Term Survivors*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley, Chichester. MR1453117

PAL, S. (2014). Likelihood-based inferential methods for some flexible cure rate models. Available at http://hdl.handle.net/11375/13688. Accessed 7 July 2021.

PAL, S. (2021). A simplified stochastic EM algorithm for cure rate model with negative binomial competing risks: An application to breast cancer data. *Stat. Med.* **40** 6387–6409. MR4339407 https://doi.org/10.1002/sim.9189

PAL, S. and ASELISEWINE, W. (2023). Supplement to "A semi-parametric promotion time cure model with support vector machine." https://doi.org/10.1214/23-AOAS1741SUPP

PAL, S. and BALAKRISHNAN, N. (2016). Destructive negative binomial cure rate model and EM-based likelihood inference under Weibull lifetime. *Statist. Probab. Lett.* **116** 9–20. MR3508515 https://doi.org/10.1016/j.spl.2016.04.005

PAL, S. and BALAKRISHNAN, N. (2017a). Likelihood inference for COM-Poisson cure rate model with interval-censored data and Weibull lifetimes. *Stat. Methods Med. Res.* **26** 2093–2113. MR3712222 https://doi.org/10.1177/0962280217708686

PAL, S. and BALAKRISHNAN, N. (2017b). Expectation maximization algorithm for Box-Cox transformation cure rate model and assessment of model misspecification under Weibull lifetimes. *IEEE J. Biomed. Health Inform.* **22** 926–934.

PAL, S. and BALAKRISHNAN, N. (2017c). An EM type estimation procedure for the destructive exponentially weighted Poisson regression cure model under generalized gamma lifetime. *J. Stat. Comput. Simul.* **87** 1107–1129. MR3606841 https://doi.org/10.1080/00949655.2016.1247843

PAL, S. and BALAKRISHNAN, N. (2017d). Likelihood inference for the destructive exponentially weighted Poisson cure rate model with Weibull lifetime and an application to melanoma data. *Comput. Statist.* **32** 429–449. MR3656969 https://doi.org/10.1007/s00180-016-0660-8

PAL, S., MAJAKWARA, J. and BALAKRISHNAN, N. (2018). An EM algorithm for the destructive COM-Poisson regression cure rate model. *Metrika* **81** 143–171. MR3756109 https://doi.org/10.1007/s00184-017-0638-8

PENG, Y. and DEAR, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56** 237–243. https://doi.org/10.1111/j.0006-341x.2000.00237.x

PENG, Y. and YU, B. (2021). *Cure Models*: *Methods*, *Applications and Implementation*. Chapman and Hall/CRC, Boca Raton.

PLATT, J. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods—Support Vector Learning* (B. Schlkopf, C. Burges and A. Smola, eds.) 185–208. MIT Press, Cambridge, MA, USA.

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

SPOONER, A., CHEN, E., SOWMYA, A., SACHDEV, P., KOCHAN, N. A., TROLLOR, J. and BRODATY, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Rep.* **10** 20410. https://doi.org/10.1038/s41598-020-77220-w

TANDON, R., ADAK, S. and KAYE, J. A. (2006). Neural networks for longitudinal studies in Alzheimer's disease. *Artif. Intell. Med.* **36** 245–255.

VAN BELLE, V., PELCKMANS, K., HUFFEL, S. V. and SUYKENS, J. A. K. (2011). Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artif. Intell. Med.* **53** 107–118. https://doi.org/10.1016/j.artmed.2011.06.006

WANG, P. and PAL, S. (2022). A two-way flexible generalized gamma transformation cure rate model. *Stat. Med.* **41** 2427–2447. MR4418444 https://doi.org/10.1002/sim.9363

XIE, Y. and YU, Z. (2021). Promotion time cure rate model with a neural network estimated nonparametric component. *Stat. Med.* **40** 3516–3532. MR4269067 https://doi.org/10.1002/sim.8980

YAKOVLEV, A. Y. and TSODIKOV, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

YIN, G. and IBRAHIM, J. G. (2005). Cure rate models: A unified approach. *Canad. J. Statist.* **33** 559–570. MR2232380 https://doi.org/10.1002/cjs.5550330407

The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2700
https://doi.org/10.1214/23-AOAS1777
Main article: https://doi.org/10.1214/18-AOAS1207

CORRIGENDUM
MODELING BIOMARKER RATIOS WITH GAMMA DISTRIBUTED COMPONENTS

BY MATTHIAS SCHMID[a]

Department of Medical Biometry, Informatics and Epidemiology, University of Bonn, [a]*matthias.schmid@imbie.uni-bonn.de*

REFERENCES

BERGER, M., WAGNER, M. and SCHMID, M. (2019). Modeling biomarker ratios with gamma distributed components. *Ann. Appl. Stat.* **13** 548–572. MR3937440 https://doi.org/10.1214/18-AOAS1207