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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

THE ANNALS OF APPLIED STATISTICS Vol. 18, No. 2, pp. 905–1772 June 2024

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*: Paul Bourgade, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012-1185, USA. Julien Dubedat, Department of Mathematics, Columbia University, New York, NY 10027, USA

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 18, Number 2, June 2024. 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
2024, Vol. 18, No. 2, 905–921
https://doi.org/10.1214/21-AOAS1586

A MARGINAL STRUCTURAL MODEL FOR PARTIAL COMPLIANCE IN SMARTS

BY WILLIAM J. ARTMAN[1,a], INDRABATI BHATTACHARYA[2,d], ASHKAN ERTEFAIE[1,b], KEVIN G. LYNCH[3,e], JAMES R. MCKAY[4,f] AND BRENT A. JOHNSON[1,c]

[1]*Department of Biostatistics and Computational Biology, University of Rochester Medical Center,*
[a]*William_Artman@URMC.Rochester.edu,* [b]*Ashkan_Ertefaie@URMC.Rochester.edu,* [c]*Brent_Johnson@URMC.Rochester.edu*
[2]*Department of Statistics, Florida State University ,* [d]*ib22g@fsu.edu*
[3]*Center for Clinical Epidemiology and Biostatistics (CCEB) and Department of Psychiatry, University of Pennsylvania,*
[e]*lynch3@pennmedicine.upenn.edu*
[4]*Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania,* [f]*james.mckay@pennmedicine.upenn.edu*

The cyclical and heterogeneous nature of many substance use disorders highlights the need to adapt the type and/or the dose of treatment to accommodate the specific and changing needs of individuals. The Adaptive Treatment for Alcohol and Cocaine Dependence study (ENGAGE) is a sequential multiple assignment randomized trial (SMART) that provided longitudinal data for constructing dynamic treatment regimes (DTRs) to improve patients' engagement in therapy. However, the high rate of noncompliance and lack of analytic tools to account for noncompliance has impeded researchers from using the data to achieve the main goal of the trial; namely, construction of individually tailored DTRs. We address this by defining our target parameter as the mean outcome under different DTRs for potential compliance strata and propose a marginal structural model with principal stratification to estimate this quantity. We model the principal strata using a Bayesian semiparametric approach. An important feature of our work is that we consider partial rather than binary compliance strata, which is more relevant in longitudinal studies. We assess the performance of our method through simulation. We illustrate its application on ENGAGE and demonstrate the optimal DTRs depend on compliance strata compared with ignoring compliance information as in intention-to-treat analyses.

REFERENCES

ANGRIST, J. D. (2006). Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology* **2** 23–44.

ANGRIST, J. D. and IMBENS, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Amer. Statist. Assoc.* **90** 431–442. MR1340501

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ARTMAN, W. J., ERTEFAIE, A., LYNCH, K. G. and MCKAY, J. R. (2020a). Bayesian set of best dynamic treatment regimes and sample size determination for SMARTs with binary outcomes. ArXiv preprint. Available at arXiv:2008.02341.

ARTMAN, W. J., NAHUM-SHANI, I., WU, T., MCKAY, J. R. and ERTEFAIE, A. (2020b). Power analysis in a SMART design: Sample size estimation for determining the best embedded dynamic treatment regime. *Biostatistics* **21** 432–448. MR4120332 https://doi.org/10.1093/biostatistics/kxy064

ARTMAN, W. J., ERTEFAIE, A., LYNCH, K. G., MCKAY, J. R. and JOHNSON, B. A. (2022). Supplement to "A marginal structural model for partial compliance in SMARTs." https://doi.org/10.1214/21-AOAS1586SUPPA, https://doi.org/10.1214/21-AOAS1586SUPPB

BARTOLUCCI, F. and GRILLI, L. (2011). Modeling partial compliance through copulas in a principal stratification framework. *J. Amer. Statist. Assoc.* **106** 469–479. MR2866975 https://doi.org/10.1198/jasa.2011.ap09094

Key words and phrases. Dynamic treatment regime, nonparametric Bayes, partial compliance, principal stratification, sequential multiple assignment randomized trial, marginal structural models.

BAYARRI, M. J. and BERGER, J. O. (2000). p values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142. MR1804239 https://doi.org/10.2307/2669749
BERRY, D. A. and HOCHBERG, Y. (1999). Bayesian perspectives on multiple comparisons. *J. Statist. Plann. Inference* **82** 215–227. MR1736444 https://doi.org/10.1016/S0378-3758(99)00044-0
BLACK, J. J. and CHUNG, T. (2014). Mechanisms of change in adolescent substance use treatment: How does treatment work? *Subst. Abuse* **35** 344–351.
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
CHAKRABORTY, B. and MURPHY, S. A. (2014). Dynamic treatment regimes. *Annu. Rev. Stat. Appl.* **1** 447–464.
CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 815–836. MR2301296 https://doi.org/10.1111/j.1467-9868.2006.00568.x
CUI, Y. and TCHETGEN TCHETGEN, E. (2021). A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *J. Amer. Statist. Assoc.* **116** 162–173. MR4227683 https://doi.org/10.1080/01621459.2020.1783272
EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.* **86** 9–17.
ERTEFAIE, A., WU, T., LYNCH, K. G. and NAHUM-SHANI, I. (2016a). Identifying a set that contains the best dynamic treatment regimes. *Biostatistics* **17** 135–148. MR3449856 https://doi.org/10.1093/biostatistics/kxv025
ERTEFAIE, A., SMALL, D., FLORY, J. and HENNESSY, S. (2016b). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *Int. J. Biostat.* **12** 219–232. MR3505695 https://doi.org/10.1515/ijb-2015-0006
ERTEFAIE, A., HSU, J. Y., PAGE, L. C. and SMALL, D. S. (2018). Discovering treatment effect heterogeneity through post-treatment variables with application to the effect of class size on mathematics scores. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 917–938. MR3832257 https://doi.org/10.1111/rssc.12265
FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379. MR1705410 https://doi.org/10.1093/biomet/86.2.365
FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x
FRANGAKIS, C. E., RUBIN, D. B. and ZHOU, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* **3** 147–164.
GASTFRIEND, D. R., FILSTEAD, W. J., REIF, S., NAJAVITS, L. M. and PARRELLA, D. P. (1995). Validity of assessing treatmen readiness in patients with substance use disorders. *Am. J. Addict.* **4** 254–260.
GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC Press/CRC, New York.
GREEVY, R., SILBER, J. H., CNAAN, A. and ROSENBAUM, P. R. (2004). Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *J. Amer. Statist. Assoc.* **99** 7–15. MR2061884 https://doi.org/10.1198/016214504000000025
GUO, M. and HEITJAN, D. F. (2010). Multiplicity-calibrated Bayesian hypothesis tests. *Biostatistics* **11** 473–483.
HERNÁN, M. Á., BRUMBACK, B. and ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 561–570.
HEWITT, C. E., TORGERSON, D. J. and MILES, J. N. V. (2006). Is there another way to take account of noncompliance in randomized controlled trials? *CMAJ, Can. Med. Assoc. J.* **175** 347. https://doi.org/10.1503/cmaj.051625
HSU, J. C. (1981). Simultaneous confidence intervals for all distances from the "best". *Ann. Statist.* **9** 1026–1034. MR0628758
HSU, J. C. (1996). *Multiple Comparisons*: *Theory and Methods*. CRC Press, London. MR1629127 https://doi.org/10.1007/978-1-4899-7180-7
JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. MR2463484 https://doi.org/10.1198/016214507000000347
KRANZLER, H. R. and MCKAY, J. R. (2012). Personalized treatment of alcohol dependence. *Curr. Psychiatry Rep.* **14** 486–493.
LABER, E. B., LIZOTTE, D. J., QIAN, M., PELHAM, W. E. and MURPHY, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.* **8** 1225–1272. MR3263118 https://doi.org/10.1214/14-EJS920
LAVORI, P. W., DAWSON, R. and RUSH, A. J. (2000). Flexible treatment strategies in chronic disease: Clinical and research implications. *Biol. Psychiatry* **48** 605–614.
LEI, H., NAHUM-SHANI, I., LYNCH, K., OSLIN, D. and MURPHY, S. A. (2012). A "SMART" design for building individualized treatment sequences. *Annu. Rev. Clin. Psychol.* **8** 21–48.

LIN, J. Y., TEN HAVE, T. R. and ELLIOTT, M. R. (2008). Longitudinal nested compliance class model in the presence of time-varying noncompliance. *J. Amer. Statist. Assoc.* **103** 462–473. MR2523985 https://doi.org/10.1198/016214507000000374

MANDEL, M. and BETENSKY, R. A. (2008). Simultaneous confidence intervals based on the percentile bootstrap approach. *Comput. Statist. Data Anal.* **52** 2158–2165. MR2418494 https://doi.org/10.1016/j.csda.2007.07.005

MARASINGHE, J. P. and AMARASINGHE, A. A. W. (2007). Noncompliance in randomized controlled trials. *CMAJ, Can. Med. Assoc. J.* **176** 1735. https://doi.org/10.1503/cmaj.1060189

MCKAY, J. R. (2009). *Treating Substance Use Disorders with Adaptive Continuing Care*. American Psychological Association, Washington.

MCKAY, J. R., LYNCH, K. G., SHEPARD, D. S., MORGENSTERN, J., FORMAN, R. F. and PETTINATI, H. M. (2005). Do patient characteristics and initial progress in treatment moderate the effectiveness of telephone-based continuing care for substance use disorders? *Addiction* **100** 216–226.

MCKAY, J. R., DRAPKIN, M. L., VAN HORN, D. H., LYNCH, K. G., OSLIN, D. W., DEPHILIPPIS, D., IVEY, M. and CACCIOLA, J. S. (2015). Effect of patient choice in an adaptive sequential randomization trial of treatment for alcohol and cocaine dependence. *J. Consult. Clin. Psychol.* **83** 1021.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. MR3937487 https://doi.org/10.1002/sim.8086

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). 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., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586 https://doi.org/10.1198/016214501753382327

NAHUM-SHANI, I., QIAN, M., ALMIRALL, D., PELHAM, W. E., GNAGY, B., FABIANO, G. A., WAXMONSKY, J. G., YU, J. and MURPHY, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychol. Methods* **17** 457.

NAHUM-SHANI, I., ALMIRALL, D., YAP, J. R. T., MCKAY, J. R., LYNCH, K. G., FREIHEIT, E. A. and DZIAK, J. J. (2020). SMART longitudinal analysis: A tutorial for using repeated outcome measures from SMART studies to compare adaptive interventions. *Psychol. Methods* **25** 1–29. https://doi.org/10.1037/met0000219

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES AND OTHERS (1992). Substance abuse and mental health services administration, office of applied studies. *Treat. Episode Data Set* **2005**.

OPSAL, A., KRISTENSEN, Ø. and CLAUSEN, T. (2019). Readiness to change among involuntarily and voluntarily admitted patients with substance use disorders. *Subst. Abuse Treat. Prev. Policy* **14** 1–10.

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *Int. J. Biostat.* **6** 8. MR2602551 https://doi.org/10.2202/1557-4679.1200

RAFTERY, D., KELLY, P. J., DEANE, F. P., BAKER, A. L., INGRAM, I., GOH, M. C., LUBMAN, D. I., CARTER, G., TURNER, A. et al. (2020). Insight in substance use disorder: A systematic review of the literature. *Addict. Behav.* 106549.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758 https://doi.org/10.1016/0270-0255(86)90088-6

ROBINS, J. M. (1999). Association, causation, and marginal structural models. *Synthese* **121** 151–179. MR1766776 https://doi.org/10.1023/A:1005285815569

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

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

ROBINS, J. M. and TSIATIS, A. A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Comm. Statist. Theory Methods* **20** 2609–2631. MR1144866 https://doi.org/10.1080/03610929108830654

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167. MR1804240 https://doi.org/10.2307/2669750

SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **106** 1331–1344. MR2896839 https://doi.org/10.1198/jasa.2011.ap10425

SIMPSON, D. D. and JOE, G. W. (1993). Motivation as a predictor of early dropout from drug abuse treatment. *Psychother. Theory Res. Pract. Train.* **30** 357.

SJÖLANDER, A. and VANSTEELANDT, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *Eur. J. Epidemiol*. **34** 809–821. https://doi.org/10.1007/s10654-019-00517-2

SLOAS, L. B., CAUDY, M. S. and TAXMAN, F. S. (2017). Is treatment readiness associated with substance use treatment engagement? An exploratory study. *J. Drug Educ*. **47** 51–67. https://doi.org/10.1177/0047237918759955

VAN HORN, D. H., DRAPKIN, M., LYNCH, K. G., RENNERT, L., GOODMAN, J. D., THOMAS, T., IVEY, M. and MCKAY, J. R. (2015). Treatment choices and subsequent attendance by substance-dependent patients who disengage from intensive outpatient treatment. *Addict. Res. Theory* **23** 391–403.

WAGENMAKERS, E.-J., LODEWYCKX, T., KURIYAL, H. and GRASMAN, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cogn. Psychol*. **60** 158–189.

WAGNER, B., RIGGS, P. and MIKULICH-GILBERTSON, S. (2015). The importance of distribution-choice in modeling substance use data: A comparison of negative binomial, beta binomial, and zero-inflated distributions. *Am. J. Drug Alcohol Abuse* **41** 489–497.

WESTFALL, P. H., JOHNSON, W. O. and UTTS, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84** 419–427. MR1467057 https://doi.org/10.1093/biomet/84.2.419

WITKIEWITZ, K., FINNEY, J. W., HARRIS, A. H., KIVLAHAN, D. R. and KRANZLER, H. R. (2015). Recommendations for the design and analysis of treatment trials for alcohol use disorders. *Alcohol. Clin. Exp. Res*. **39** 1557–1570.

WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

WOOLDRIDGE, J. M. (2015). *Introductory Econometrics*: *A Modern Approach*. Nelson Education, Cincinnati, OH: South-Western.

ZEMORE, S. E., WARE, O. D., GILBERT, P. A. and PINEDO, M. (2021). Barriers to retention in substance use treatment: Validation of a new, theory-based scale. *J. Subst. Abuse Treat*. **131** 108422. https://doi.org/10.1016/j.jsat.2021.108422

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 922–945
https://doi.org/10.1214/22-AOAS1702

A HIGH-DIMENSIONAL APPROACH TO MEASURE CONNECTIVITY IN THE FINANCIAL SECTOR

BY SUMANTA BASU[1,a], SREYOSHI DAS[1,b], GEORGE MICHAILIDIS[2,c] AND AMIYATOSH PURNANANDAM[3,d]

[1]*Department of Statistics and Data Science, Cornell University,* [a]*sumbose@cornell.edu,* [b]*sreyoshi.das@cornell.edu*
[2]*Department of Statistics and Informatics Institute, University of Florida,* [c]*gmichail@ufl.edu*
[3]*Ross School of Business, University of Michigan,* [d]*amiyatos@umich.edu*

Data-driven network models to measure systemic risk in the financial sector and identify "too-connected-to-fail" institutions are becoming increasingly common in financial applications. Existing statistical methods for building such networks either take a pairwise approach of fitting many bivariate models or a system-wide approach of fitting penalized regression models. The former strategy is prone to large false positive selection, while the latter suffers from shrinkage bias and lack of formal inference machinery. These issues are accentuated in small sample, low signal-to-noise settings common in financial data. Building up on recent advances in high-dimensional inference, we propose debiased lasso Penalized Vector Autoregression (DLVAR), a method for building financial networks that addresses these limitations. Our empirical analysis highlights the importance of debiasing in a way that increases power of the algorithm in finite samples. We also provide formal inference guarantees of Granger causality tests in high-dimension to justify our method. We apply DLVAR to the stock returns of U.S. large financial institutions covering the period 1990–2021 and illustrate its usefulness in detecting systemically risky periods and institutions, especially during the Great Financial Crisis of 2008–2009 and the most recent Covid-19 related market shock.

REFERENCES

ACHARYA, V. V., PEDERSEN, L. H., PHILIPPON, T. and RICHARDSON, M. (2017). Measuring systemic risk. *Rev. Financ. Stud.* **30** 2–47.

AHELEGBEY, D. F., BILLIO, M. and CASARIN, R. (2016a). Sparse graphical vector autoregression: A Bayesian approach. *Ann. Econ. Statist.* **123/124** 333–361.

AHELEGBEY, D. F., BILLIO, M. and CASARIN, R. (2016b). Bayesian graphical models for structural vector autoregressive processes. *J. Appl. Econometrics* **31** 357–386. MR3481367 https://doi.org/10.1002/jae.2443

ALLEN, F. and GALE, D. (2000). Financial contagion. *J. Polit. Econ.* **108** 1–33.

ASHCRAFT, A. B., GOLDSMITH-PINKHAM, P. and VICKERY, J. I. (2010). MBS ratings and the mortgage credit boom.

BARIGOZZI, M. and BROWNLEES, C. (2019). NETS: Network estimation for time series. *J. Appl. Econometrics* **34** 347–364. MR3948470 https://doi.org/10.1002/jae.2676

BASU, S., DAS, S., MICHAILIDIS, G. and PURNANANDAM, A. (2024). Supplement to "A high-dimensional approach to measure connectivity in the financial sector." https://doi.org/10.1214/22-AOAS1702SUPP

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

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

BILLIO, M., GETMANSKY, M. and LO, A. W. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *J. Financ. Econ.* **104** 535–559.

BLUHM, M. and KRAHNEN, P. J. (2014). Systemic risk in an interconnected banking system with endogenous asset markets. *J. Financ. Stab.* **13** 75–94.

Key words and phrases. Debiased lasso, systemic risk, vector autoregression.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

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

CALOMIRIS, C. W. and KHAN, U. (2015). An assessment of TARP assistance to financial institutions. *J. Econ. Perspect.* **29** 53–80.

CORDEIRO, G. M. and MCCULLAGH, P. (1991). Bias correction in generalized linear models. *J. Roy. Statist. Soc. Ser. B* **53** 629–643. MR1125720

CORDEIRO, G. M. and VASCONCELLOS, K. L. P. (1997). Bias correction for a class of multivariate nonlinear regression models. *Statist. Probab. Lett.* **35** 155–164. MR1483269 https://doi.org/10.1016/S0167-7152(97)00009-6

DEMIRER, M., DIEBOLD, F. X., LIU, L. and YILMAZ, K. (2018). Estimating global bank network connectedness. *J. Appl. Econometrics* **33** 1–15. MR3771571 https://doi.org/10.1002/jae.2585

DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and R-software `hdi`. *Statist. Sci.* **30** 533–558. MR3432840 https://doi.org/10.1214/15-STS527

DIEBOLD, F. X. and YILMAZ, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econometrics* **182** 119–134. MR3212765 https://doi.org/10.1016/j.jeconom.2014.04.012

ENGLE, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econom. Statist.* **20** 339–350. MR1939905 https://doi.org/10.1198/073500102288618487

FSOC Financial Stability Oversight Council (2019). 2019 Annual Report. available at: https://home.treasury.gov/system/files/261/FSOC2019AnnualReport.pdf.

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

KRAMPE, J., KREISS, J.-P. and PAPARODITIS, E. (2021). Bootstrap based inference for sparse high-dimensional time series models. *Bernoulli* **27** 1441–1466. MR4278792 https://doi.org/10.3150/20-bej1239

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 129–156. Univ. California Press, Berkeley-Los Angeles, CA. MR0084918

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038 https://doi.org/10.1214/12-AOS1018

LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

MISHKIN, F. S. (2011). Over the cliff: From the subprime to the global financial crisis. *J. Econ. Perspect.* **25** 49–70.

PHILIPPON, T. and SCHNABL, P. (2013). Efficient recapitalization. *J. Finance* **68** 1–42.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166 https://doi.org/10.1093/biomet/ass043

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHENG, L. and RASKUTTI, G. (2019). Testing for high-dimensional network parameters in auto-regressive models. *Electron. J. Stat.* **13** 4977–5043. MR4041701 https://doi.org/10.1214/19-EJS1646

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 946–965
https://doi.org/10.1214/23-AOAS1811

BAYESIAN HIERARCHICAL MODELLING OF SPARSE COUNT PROCESSES IN RETAIL ANALYTICS

BY JAMES PITKIN[2,a], IOANNA MANOLOPOULOU[2,1,b] AND GORDON ROSS[3,c]

[1]*The Alan Turing Institute*
[2]*University College London,* [a]*james.pitkin@cantab.net,* [b]*i.manolopoulou@ucl.ac.uk*
[3]*University of Edinburgh,* [c]*gordon.ross@ed.ac.uk*

The field of retail analytics has been transformed by the availability of rich data, which can be used to perform tasks such as demand forecasting and inventory management. However, one task which has proved more challenging is the forecasting of demand for products which exhibit very few sales. The sparsity of the resulting data limits the degree to which traditional analytics can be deployed. To combat this, we represent sales data as a structured sparse multivariate point process, which allows for features such as autocorrelation, cross-correlation, and temporal clustering, known to be present in sparse sales data. We introduce a Bayesian point process model to capture these phenomena, which includes a hurdle component to cope with sparsity and an exciting component to cope with temporal clustering within and across products. We then cast this model within a Bayesian hierarchical framework, to allow the borrowing of information across different products, which is key in addressing the data sparsity per product. We conduct a detailed analysis, using real sales data, to show that this model outperforms existing methods in terms of predictive power, and we discuss the interpretation of the inference.

REFERENCES

BERRY, L. R., HELMAN, P. and WEST, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction–sales time series. *Int. J. Forecast.* **36** 552–569. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2019.07.007

BLUNDELL, C., BECK, J. and HELLER, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems* 2600–2608.

CHAPADOS, N. (2014). Effective Bayesian modeling of groups of related count time series. arXiv preprint. Available at arXiv:1405.3738.

DO CROSTON, J. (1972). Forecasting and stock control for intermittent demands. *Oper. Res. Q.* 289–303.

GARDNER, G. S. (2006). Exponential smoothing: The state of the art—part II. *Int. J. Forecast.* **22** 637–666.

FERREIRA, K. J., LEE, B. H. A. and SIMCHI-LEVI, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. *Manuf. Serv. Oper. Manag.*

GHOBBAR, A. A. and FRIEND, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Comput. Oper. Res.* **30** 2097–2114.

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83

HEIDELBERGER, P. and WELCH, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM* **24** 233–245. MR0611745 https://doi.org/10.1145/358598.358630

KOURENTZES, N. (2013). Intermittent demand forecasts with neural networks. *Int. J. Prod. Econ.* **143** 198–206.

LAI, E. L., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2016). Topic time series analysis of microblogs. *IMA J. Appl. Math.* **81** 409–431. MR3564661 https://doi.org/10.1093/imamat/hxw025

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

Key words and phrases. Self-excitation, cross-excitation, hurdle model, Hawkes process, intermittent demand, slow-moving-inventory, demand forecasting.

SEEGER, M. W., SALINAS, D. and FLUNKERT, V. (2016a). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds.) **29**. Curran Associates, Red Hook.

SEEGER, M. W., SALINAS, D. and FLUNKERT, V. (2016b). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* 4646–4654.

MISHRA, P., YUAN, X.-M., HUANG, G. and DUC, T. T. H. (2014). *Intermittent Demand Forecast*: *Robustness Assessment for Group Method of Data Handling*.

MULLAHY, J. (1986). Specification and testing of some modified count data models. *J. Econometrics* **33** 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

NIETO-BARAJAS, L. E. and CONTRERAS-CRISTÁN, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Anal.* **9** 147–169. MR3188303 https://doi.org/10.1214/13-BA852

PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531 https://doi.org/10.1214/11-AOAS513

POUR, A. N., TABAR, B. R. and RAHIMZADEH, A. (2008). A hybrid neural network and traditional approach for forecasting lumpy demand. *Proc. World Acad. Sci. Eng. Technol.* **30** 384–389.

RANGAPURAM, S. S., SEEGER, M. W., GASTHAUS, J., STELLA, L., WANG, Y. and JANUSCHOWSKI, T. (2018). Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) **31**. Curran Associates, Red Hook.

SAHU, S. K., BAFFOUR, B., HARPER, P. R., MINTY, J. H. and SARRAN, C. (2014). A hierarchical Bayesian model for improving short-term forecasting of hospital demand by including meteorological information. *J. Roy. Statist. Soc. Ser. A* **177** 39–61. MR3158666 https://doi.org/10.1111/rssa.12008

SALINAS, D., FLUNKERT, V., GASTHAUS, J. and DEEPAR, T. J. (2020). Probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* **36** 1181–1191. ISSN 0169-2070.

SHENSTONE, L. and HYNDMAN, R. J. (2005). Stochastic models underlying Croston's method for intermittent demand forecasting. *J. Forecast.* **24** 389–402. MR2206931 https://doi.org/10.1002/for.963

SNYDER, R. D., ORD, J. K. and BEAUMONT, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *Int. J. Forecast.* **28** 485–496.

TADESSE, M. G. and VANNUCCI, M. (2021). *Handbook of Bayesian Variable Selection*. CRC Press, Boca Raton.

ZHOU, K., ZHA, H. and SONG, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *AISTATS* **13** 641–649.

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

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 966–990
https://doi.org/10.1214/23-AOAS1816
This research was funded, in whole or in part, by EPSRC, EP/V049968/1.

A LATENT MIXTURE MODEL FOR HETEROGENEOUS CAUSAL MECHANISMS IN MENDELIAN RANDOMIZATION

BY DANIEL IONG[1,a], QINGYUAN ZHAO[2,c] AND YANG CHEN[1,b]

[1]*Department of Statistics, University of Michigan,* [a]*daniong@umich.edu,* [b]*ychenang@umich.edu*
[2]*Statistical Laboratory, University of Cambridge,* [c]*qyzhao@statslab.cam.ac.uk*

Mendelian randomization (MR) is a popular method in epidemiology and genetics that uses genetic variation as instrumental variables for causal inference. Existing MR methods usually assume most genetic variants are valid instrumental variables that identify a common causal effect. There is a general lack of awareness that this effect homogeneity assumption can be violated when there are multiple causal pathways involved, even if all the instrumental variables are valid. In this article we introduce a latent mixture model MR-Path that groups instruments that yield similar causal effect estimates together. We develop a Monte Carlo EM algorithm to fit this mixture model, derive approximate confidence intervals for uncertainty quantification, and adopt a modified Bayesian Information Criterion (BIC) for model selection. We verify the efficacy of the Monte Carlo EM algorithm, confidence intervals, and model selection criterion using numerical simulations. We identify potential mechanistic heterogeneity when applying our method to estimate the effect of high-density lipoprotein cholesterol on coronary heart disease and the effect of adiposity on type II diabetes.

REFERENCES

AKIYAMA, M., OKADA, Y., KANAI, M., TAKAHASHI, A., MOMOZAWA, Y., IKEDA, M., IWATA, N., IKEGAWA, S., HIRATA, M. et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49** 1458.

ANDERSON, T. W. and RUBIN, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.* **20** 46–63. MR0028546 https://doi.org/10.1214/aoms/1177730090

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ARMITAGE, J., HOLMES, M. V. and PREISS, D. (2019). Cholesteryl ester transfer protein inhibition for preventing cardiovascular events: JACC review topic of the week. *J. Am. Coll. Cardiol.* **73** 477–487. https://doi.org/10.1016/j.jacc.2018.10.072

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

BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525. https://doi.org/10.1093/ije/dyv080

BOWDEN, J., DEL GRECO M, F., MINELLI, C., DAVEY SMITH, G., SHEEHAN, N. and THOMPSON, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36** 1783–1802. MR3648622 https://doi.org/10.1002/sim.7221

BOYLE, E. A., LI, Y. I. and PRITCHARD, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169** 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

BUNIELLO, A., MACARTHUR, J. A. L., CEREZO, M., HARRIS, L. W. and HAYHURST, J. MALANGONE, C. MCMAHON, A. MORALES, J. MOUNTJOY, E. et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** D1005–D1012.

BURGESS, S., BOWDEN, J. and FALL, T. (2017). Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28** 30–42. https://doi.org/10.1097/EDE.0000000000000559

Key words and phrases. Causal inference, instrumental variables, EM algorithm, Monte Carlo sampling, HDL cholesterol, diabetes.

BURGESS, S., BUTTERWORTH, A. and THOMPSON, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37** 658–665. https://doi.org/10.1002/gepi.21758

BURGESS, S., FOLEY, C. N., ALLARA, A., STALEY, J. R. and HOWSON, J. M. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat. Commun.* **11** 376. https://doi.org/10.1038/s41467-019-14156-4

CAFFO, B. S., JANK, W. and JONES, G. L. (2005). Ascent-based Monte Carlo expectation-maximization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 235–251. MR2137323 https://doi.org/10.1111/j.1467-9868.2005.00499.x

DAVEY SMITH, G. and PHILLIPS, A. N. (2020). Correlation without a cause: An epidemiological odyssey. *Int. J. Epidemiol.* **49** 4–14. https://doi.org/10.1093/ije/dyaa016

DIDELEZ, V. and SHEEHAN, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* **16** 309–330. MR2395652 https://doi.org/10.1177/0962280206077743

FOLEY, C. N., KIRK, P. D. W. and BURGESS, S. (2019). MR-Clust: Clustering of genetic variants in Mendelian randomization with similar causal estimates. *Bioinformatics*. https://doi.org/10.1101/2019.12.18.881326

IBRAHIM, J. G., ZHU, H. and TANG, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *J. Amer. Statist. Assoc.* **103** 1648–1658. MR2510293 https://doi.org/10.1198/016214508000001057

IONG, D., ZHAO, Q. and CHEN, Y. (2024). Supplement to "A Latent Mixture Model for Heterogeneous Causal Mechanisms in Mendelian Randomization." https://doi.org/10.1214/23-AOAS1816SUPP

JI, Y., YIORKAS, A. M., FRAU, F., MOOK-KANAMORI, D., STAIGER, H., THOMAS, E. L., ATABAKI-PASDAR, N., CAMPBELL, A., TYRRELL, J. et al. (2019). Genome-wide and abdominal MRI data provide evidence that a genetically determined favorable adiposity phenotype is characterized by lower ectopic liver fat and lower risk of type 2 diabetes, heart disease, and hypertension. *Diabetes* **68** 207–219.

KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. MR3494648 https://doi.org/10.1080/01621459.2014.994705

KETTUNEN, J., DEMIRKAN, A., WÜRTZ, P., DRAISMA, H. H. M., HALLER, T., RAWAL, R., VAARHORST, A., KANGAS, A., LYYTIKÄINEN, L.-P. et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7** 1–9.

LI, K.-H. (2004). The sampling/importance resampling algorithm. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat.* 265–276. Wiley, Chichester. MR2138262 https://doi.org/10.1002/0470090456.ch24

LIU, X., LI, Y. I. and PRITCHARD, J. K. (2019). Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177** 1022–1034.

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E., PERS, T. H., DAY, F. R., POWELL, C., VEDANTAM, S., BUCHKOVICH, M. L. et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206. https://doi.org/10.1038/nature14177

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** 226–233. MR0676213

MAHAJAN, A., TALIUN, D., THURNER, M., ROBERTSON, N. R., TORRES, J. M. RAYNER, N. W. PAYNE, A. J. STEINTHORSDOTTIR, V. SCOTT, R. A. et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50** 1505–1513.

NEATH, R. C. (2013). On convergence properties of the Monte Carlo EM algorithm. In *Advances in Modern Statistical Theory and Applications*: *A Festschrift in Honor of Morris L. Eaton*. *Inst. Math. Stat.* (*IMS*) *Collect.* **10** 43–62. IMS, Beachwood, OH. MR3586938

NIKPAY, M., GOEL, A., WON, H.-H., HALL, L. M., WILLENBORG, C., KANONI, S., SALEHEEN, D., KYRIAKOU, T., NELSON, C. P. et al. (2015). A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47** 1121–1130. https://doi.org/10.1038/ng.3396

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

QI, G. and CHATTERJEE, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat. Commun.* **10**. 1941. https://doi.org/10.1038/s41467-019-09432-2

RADER, D. J., HOVINGH, K. G. (2014). HDL and cardiovascular disease. *Lancet* **384** 618–625.

SHAPLAND, C. Y., ZHAO, Q. and BOWDEN, J. (2022). Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy. *Stat. Med.* **41** 1100–1119. MR4389988 https://doi.org/10.1002/sim.9320

SMITH, G. D. and EBRAHIM, S. (2004). Mendelian randomization: Prospects, potentials, and limitations. *Int. J. Epidemiol.* **33** 30–42. https://doi.org/10.1093/ije/dyh132

TESLOVICH, T. M., MUSUNURU, K., SMITH, A. V., EDMONDSON, A. C., STYLIANOU, I. M., KOSEKI, M., PIRRUCCELLO, J. P., RIPATTI, S., CHASMAN, D. I. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** 707–713.

TOKDAR, S. T. and KASS, R. E. (2010). Importance sampling: A review. *Wiley Interdiscip. Rev.: Comput. Stat.* **2** 54–60. https://doi.org/10.1002/wics.56

VERBANCK, M., CHEN, C.-Y., NEALE, B. and DO, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50** 693–698. https://doi.org/10.1038/s41588-018-0099-7

VOIGHT, B. F., PELOSO, G. M., ORHO-MELANDER, M., FRIKKE-SCHMIDT, R., BARBALIC, M., JENSEN, M. K., HINDY, G., HÓLM, H., DING, E. L. et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: A Mendelian randomisation study. *Lancet* (*London, England*) **380** 572–80. https://doi.org/10.1016/S0140-6736(12)60312-2

WANG, J., ZHAO, Q., BOWDEN, J., HEMANI, G., SMITH, G. D., SMALL, D. S. and ZHANG, N. R. (2020). Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. https://doi.org/10.1101/2020.05.06.077982.

WILLER, C. J., SCHMIDT, E. M., SENGUPTA, S., PELOSO, G. M., GUSTAFSSON, S., KANONI, S., GANNA, A., CHEN, J., BUCHKOVICH, M. L. et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45** 1274–1283. https://doi.org/10.1038/ng.2797

WOOD, A. R., JONSSON, A., JACKSON, A. U., WANG, N., VAN LEEWEN, N., PALMER, N. D., KOBES, S., DEELEN, J., BOQUETE-VILARINO, L. et al. (2017). A genome-wide association study of IVGTT-based measures of first-phase insulin secretion refines the underlying physiology of type 2 diabetes variants. *Diabetes* **66** 2296–2309.

WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867 https://doi.org/10.1214/aos/1176346060

ZHAO, Q., WANG, J., HEMANI, G., BOWDEN, J. and SMALL, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Statist.* **48** 1742–1769. MR4124342 https://doi.org/10.1214/19-AOS1866

ZHAO, Q., WANG, J., MIAO, Z., ZHANG, N., HENNESSY, S. and SMALL, D. S. (2019). The role of lipoprotein subfractions in coronary artery disease: A Mendelian randomization study. bioRxiv 691089. https://doi.org/10.1101/691089

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 991–1009
https://doi.org/10.1214/23-AOAS1819

IDENTIFICATION OF INFLUENCING FACTORS ON SELF-REPORTED COUNT DATA WITH MULTIPLE POTENTIAL INFLATED VALUES

BY YANG LI[1,a], MINGCONG WU[1,b], MENGYUN WU[2,c] AND SHUANGGE MA[3,d]

[1]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [a]*yang.li@ruc.edu.cn,* [b]*wumingcong@ruc.edu.cn*

[2]*School of Statistics and Management, Shanghai University of Finance and Economics,* [c]*wu.mengyun@mail.shufe.edu.cn*

[3]*Department of Biostatistics, Yale School of Public Health,* [d]*shuangge.ma@yale.edu*

The online chauffeured service demand (OCSD) research is an exploratory market study of designated driver services in China. Researchers are interested in the influencing factors of chauffeured service adoption and usage and have collected relevant data using a self-reported questionnaire. As self-reported count measure data is typically inflated, there exist challenges to its validity, which may bias estimation and increase error in empirical research. Motivated by the analysis of self-reported data with multiple inflated values, we propose a novel approach to simultaneously achieve data-driven inflated value selection and identification of important influencing factors. In particular, the regularization technique is applied to the mixing proportions of inflated values and the regression parameters to obtain shrinkage estimates. We analyze the OCSD data with the proposed approach, deriving insights into the determinants impacting service demand. The proper interpretations and implications contribute to service promotion and related policy optimization. Extensive simulation studies and consistent asymptotic properties further establish the effectiveness of the proposed approach.

REFERENCES

BANERJEE, P., GARAI, B., MALLICK, H., CHOWDHURY, S. and CHATTERJEE, S. (2018). A note on the adaptive LASSO for zero-inflated Poisson regression. *J. Probab. Stat.* 2834183, 9. MR3898741 https://doi.org/10.1155/2018/2834183

BOCCI, C., GRASSINI, L. and ROCCO, E. (2021). A multiple inflated negative binomial hurdle regression model: Analysis of the Italians' tourism behaviour during the Great Recession. *Stat. Methods Appl.* **30** 1109–1133. MR4324404 https://doi.org/10.1007/s10260-020-00542-6

BUU, A., JOHNSON, N. J., LI, R. and TAN, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Stat. Med.* **30** 2326–2340. MR2830011 https://doi.org/10.1002/sim.4268

CAI, T., XIA, Y. and ZHOU, Y. (2021). Generalized inflated discrete models: A strategy to work with multimodal discrete distributions. *Sociol. Methods Res.* **50** 365–400. MR4198562 https://doi.org/10.1177/0049124118782535

CAI, Z. and WANG, X. (2014). Selection of mixed copula model via penalized likelihood. *J. Amer. Statist. Assoc.* **109** 788–801. MR3223750 https://doi.org/10.1080/01621459.2013.873366

CHEN, C.-S. and SHEN, C.-W. (2022). Distribution-free model selection for longitudinal zero-inflated count data with missing responses and covariates. *Stat. Med.* **41** 3180–3198. MR4444893 https://doi.org/10.1002/sim.9411

CHEN, T., WU, P., TANG, W., ZHANG, H., FENG, C., KOWALSKI, J. and TU, X. M. (2016). Variable selection for distribution-free models for longitudinal zero-inflated count responses. *Stat. Med.* **35** 2770–2785. MR3513717 https://doi.org/10.1002/sim.6892

CRAWFORD, F. W., WEISS, R. E. and SUCHARD, M. A. (2015). Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. *Ann. Appl. Stat.* **9** 572–596. MR3371326 https://doi.org/10.1214/15-AOAS809

DUBRAY, S., GÉRARD, M., BEAULIEU-PRÉVOST, D. and COURTOIS, F. (2017). Validation of a self-report questionnaire assessing the bodily and physiological sensations of orgasm. *J. Sex. Med.* **14** 255–263.

Key words and phrases. Inflated value selection, investigation, negative binomial data modeling, penalization.

ERSCHE, K. D., LIM, T.-V., WARD, L. H., ROBBINS, T. W. and STOCHL, J. (2017). Creature of habit: A self-report measure of habitual routines and automatic tendencies in everyday life. *Pers. Individ. Differ.* **116** 73–85.

GARAY, A. M., HASHIMOTO, E. M., ORTEGA, E. M. M. and LACHOS, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Comput. Statist. Data Anal.* **55** 1304–1318. MR2741416 https://doi.org/10.1016/j.csda.2010.09.019

GILES, D. (2007). Modeling inflated count data. In *MODSIM*07*—Land, Water and Environmental Management: Integrated Systems for Sustainability, Proceedings*.

GILES, D. (2010). Hermite regression analysis of multi-modal count data. *Econ. Bull.* **30** 2936–2945.

HANSEN, B. (2015). Punishment and deterrence: Evidence from drunk driving. *Amer. Econ. Rev.* **105** 1581–1617.

HOPP, T., FERRUCCI, P. and VARGO, C. J. (2020). Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data–based analysis of countermedia content dissemination on Facebook and Twitter. *Hum. Commun. Res.* **46** 357–384.

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662 https://doi.org/10.1198/016214507000000590

KHALILI, A. and LIN, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics* **69** 436–446. MR3071062 https://doi.org/10.1111/biom.12020

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

LI, Q., TSO, G. K. F., QIN, Y., LOVEJOY, T. I., HECKMAN, T. G. and LI, Y. (2019). Penalized multiple inflated values selection method with application to SAFER data. *Stat. Methods Med. Res.* **28** 3205–3225. MR4002694 https://doi.org/10.1177/0962280218797148

LI, Y., WU, M., WU, M. and MA, S. (2024). Supplement to "Identification of influencing factors on self-reported count data with multiple potential inflated values." https://doi.org/10.1214/23-AOAS1819SUPPA, https://doi.org/10.1214/23-AOAS1819SUPPB

SHARMA, P., CHEN, I. S. I. and LUK, S. T. K. (2012). Gender and age as moderators in the service evaluation process. *J. Serv. Mark.* **26** 102–114.

SU, X., FAN, J., LEVINE, R. A., TAN, X. and TRIPATHI, A. (2013). Multiple-inflation Poisson model with L_1 regularization. *Statist. Sinica* **23** 1071–1090. MR3114705

TABRIZI, E., BAHRAMI SAMANI, E. and GANJALI, M. (2020). Identifiability of parameters in longitudinal correlated Poisson and inflated beta regression model with non-ignorable missing mechanism. *Statistics* **54** 524–543. MR4100722 https://doi.org/10.1080/02331888.2020.1748883

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

WANG, H. and HEITJAN, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Stat. Med.* **27** 3789–3804. MR2526609 https://doi.org/10.1002/sim.3281

WANG, Z., MA, S. and WANG, C.-Y. (2015). Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biom. J.* **57** 867–884. MR3394815 https://doi.org/10.1002/bimj.201400143

XIE, F.-C., LIN, J.-G. and WEI, B.-C. (2014). Bayesian zero-inflated generalized Poisson regression model: Estimation and case influence diagnostics. *J. Appl. Stat.* **41** 1383–1392. MR3268901 https://doi.org/10.1080/02664763.2013.871508

XIE, Y., XU, L., DENG, X., HONG, Y., KOLIVRAS, K. and GAINES, D. N. (2019). Spatial variable selection and an application to Virginia Lyme disease emergence. *J. Amer. Statist. Assoc.* **114** 1466–1480. MR4047274 https://doi.org/10.1080/01621459.2018.1564670

YEE, T. W. and MA, C. (2022). Generally-altered,-inflated,-truncated and-deflated regression, with application to heaped and seeped Data. Preprint. Available at arXiv:2208.12972.

ZENG, P., WEI, Y., ZHAO, Y., LIU, J., LIU, L., ZHANG, R., GOU, J., HUANG, S. and CHEN, F. (2014). Variable selection approach for zero-inflated count data via adaptive lasso. *J. Appl. Stat.* **41** 879–894. MR3291791 https://doi.org/10.1080/02664763.2013.858672

ZHONG, T., ZHANG, Q., HUANG, J., WU, M. and MA, S. (2023). Heterogeneity analysis via integrating multi-sources high-dimensional data with applications to cancer studies. *Statist. Sinica* **33** 729–758. MR4575322 https://doi.org/10.5705/ss.202021.0002

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
2024, Vol. 18, No. 2, 1010–1034
https://doi.org/10.1214/23-AOAS1820

READABILITY PREDICTION: HOW MANY FEATURES ARE NECESSARY?

BY FLORIAN SCHWENDINGER[1,a], LAURA VANA[2,b] AND KURT HORNIK[3,c]

[1]*Department of Statistics, University of Klagenfurt,* [a]*FlorianSchwendinger@gmx.at*
[2]*Institute of Statistics and Mathematical Methods in Economics, TU Wien,* [b]*Laura.Vana@tuwien.ac.at*
[3]*Institute for Statistics and Mathematics, Wirtschaftsuniversität Wien,* [c]*Kurt.Hornik@wu.ac.at*

Traditionally, readability prediction has relied on readability formulas, which are based on shallow text characteristics such as average word and sentence length. With recent advances in text mining and natural language processing, more complex text properties can be incorporated into readability prediction models, with papers in the literature suggesting to use up to 200 features for predicting text readability. However, many of the features generated using natural language processing tools are highly correlated and can be thought to measure similar latent text properties. When dealing with a high-dimensional space of correlated features, removing the redundant variables has two advantages: (1) improving interpretability and (2) increasing the predictive power of the model. In this paper we propose an ordinal version of the averaged lasso, which combines hierarchical clustering with the lasso, in order to identify relevant features for readability prediction. We illustrate the approach on two corpora and show improved prediction accuracy when benchmarking against a set of competing models. The annotated corpora as well as the steps necessary for feature creation are freely available as R packages, thus allowing the obtained results to be directly incorporated into a readability estimation pipeline.

REFERENCES

AGRESTI, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2742515 https://doi.org/10.1002/9780470594001

AIROLDI, E. M. and BISCHOF, J. M. (2016). Improving and evaluating topic models and other models of text. *J. Amer. Statist. Assoc.* **111** 1381–1403. MR3601693 https://doi.org/10.1080/01621459.2015.1051182

BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436 https://doi.org/10.1198/016214505000000628

BARTLETT, M. S. (1937). The statistical conception of mental factors. *Br. J. Psychol. Gen. Sect.* **28** 97–104.

BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825 https://doi.org/10.1111/j.1541-0420.2007.00843.x

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32. MR3874153

BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454 https://doi.org/10.1214/07-STS242

BÜHLMANN, P., RÜTIMANN, P., VAN DE GEER, S. and ZHANG, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *J. Statist. Plann. Inference* **143** 1835–1858. MR3095072 https://doi.org/10.1016/j.jspi.2013.05.019

CHALL, J. S. and DALE, E. (1995). *Readability Revisited*: *The New Dale–Chall Readability Formula*. Brookline Books, Brookline.

CHEN, T., HE, T., BENESTY, M., KHOTILOVICH, V., TANG, Y., CHO, H., CHEN, K., MITCHELL, R., CANO, I., et al. (2020). *xgboost*: Extreme gradient boosting R package version 1.0.0.2.

CHRISTENSEN, R. H. B. (2019). *ordinal*—Regression models for ordinal data R package version 2019.12-10.

CROSSLEY, S. A., SKALICKY, S., DASCALU, M., MCNAMARA, D. S. and KYLE, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Process.* **54** 340–359. https://doi.org/10.1080/0163853x.2017.1296264

DALE, E. and CHALL, J. S. (1948). A formula for predicting readability. *Educ. Res. Bull.* **27** 11–28.

Key words and phrases. Averaged ordinal lasso, model selection, NLP, pipeline, readability prediction.

DALE, E. and CHALL, J. S. (1949). The concept of readability. *Elem. Engl.* **26** 19–26.
DE CLERCQ, O. and HOSTE, V. (2016). All mixed up? Finding the optimal feature set of general readability prediction and its application to English and Dutch. *Comput. Linguist.* **42** 457–490. MR3553984 https://doi.org/10.1162/COLI_a_00255
DE CLERCQ, O., HOSTE, V., DESMET, B., VAN OOSTEN, P., DE COCK, M. and MACKEN, L. (2014). Using the crowd for readability prediction. *Nat. Lang. Eng.* **20** 293–325. https://doi.org/10.1017/s1351324912000344
DEVLIN, J., CHANG, M.-W., LEE, K. and TOUTANOVA, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.
DUBAY, W. H. (2004). The principles of readability. Technical report, online submission.
DUTTA, S. and DAI, F. (2021). Fad: Factor analysis for data R package version 0.3-3.
FENG, L., ELHADAD, N. and HUENERFAUTH, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the* 12*th Conference of the European Chapter of the Association for Computational Linguistics. EACL* '09 229–237. Association for Computational Linguistics, Stroudsburg, PA, USA.
FENG, L., JANSCHE, M., HUENERFAUTH, M. and ELHADAD, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the* 23*rd International Conference on Computational Linguistics: Posters. COLING* **10** 276–284. Association for Computational Linguistics, Stroudsburg, PA, USA.
FLESCH, R. (1948). A new readability yardstick. *J. Appl. Psychol.* **32** 221–233. https://doi.org/10.1037/h0057532
FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
GHOSE, A. and IPEIROTIS, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowl. Data Eng.* **23** 1498–1512. https://doi.org/10.1109/tkde.2010.188
GUNNING, R. (1952). *The Technique of Clear Writing*. McGraw-Hill, New York.
HEILMAN, M., COLLINS-THOMPSON, K., CALLAN, J. and ESKENAZI, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies* 2007: *The Conference of the North American Chapter of the Association for Computational Linguistics*; *Proceedings of the Main Conference* 460–467.
HORNUNG, R. (2019a). Ordinal forests. *J. Classification* **37** 4–17. MR4111881 https://doi.org/10.1007/s00357-018-9302-x
HORNUNG, R. (2019b). *ordinalForest*: Ordinal forests: Prediction and variable ranking with ordinal target variables R package version 2.3-1.
HOTHORN, T. and BÜHLMANN, P. (2006). Model-based boosting in high dimensions. *Bioinformatics* **22** 2828–2829. https://doi.org/10.1093/bioinformatics/btl462
HOTHORN, T., BÜHLMANN, P., DUDOIT, S. and MOLINARO, A. (2006). Survival ensembles. *Biostatistics* **7** 355–373. https://doi.org/10.1093/biostatistics/kxj011
HOTHORN, T., BÜHLMANN, P., KNEIB, T., SCHMID, M. and HOFNER, B. (2010). Model-based boosting 2.0. *J. Mach. Learn. Res.* **11** 2109–2113. MR2719848
HOTHORN, T., BÜHLMANN, P., KNEIB, T., SCHMID, M. and HOFNER, B. (2020). *mboost*: Model-based boosting R package version 2.9-2.
HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267 https://doi.org/10.1198/106186006X133933
HOTHORN, T. and ZEILEIS, A. (2015). *partykit*: A modular toolkit for recursive partytioning in R. *J. Mach. Learn. Res.* **16** 3905–3909. MR3450556
HU, N., BOSE, I., KOH, N. S. and LIU, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decis. Support Syst.* **52** 674–684. https://doi.org/10.1016/j.dss.2011.11.002
ISLAM, M. Z. (2015). Multilingual text classification using information-theoretic features. Ph.D. thesis, Dept. Computer Science.
JURAFSKY, D. and MARTIN, J. H. (2009). *Speech and Language Processing*, 2nd ed. Prentice Hall, USA.
KATE, R., LUO, X., PATWARDHAN, S., FRANZ, M., FLORIAN, R., MOONEY, R. and ROUKOS, S. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the* 23*rd International Conference on Computational Linguistics* 546–554.
KIM, J. Y., COLLINS-THOMPSON, K., BENNETT, P. N. and DUMAIS, S. T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* 213–222. https://doi.org/10.1145/2124295.2124323
KINCAID, J. P., FISHBURNE JR., R. P., ROGERS, R. L. and CHISSOM, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted. Personnel technical report, Naval Technical Training Command Millington TN Research Branch.
LEHAVY, R., LI, F. and MERKLEY, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Account. Rev.* **86** 1087–1115.

LEROY, G., HELMREICH, S., COWIE, J. R., MILLER, T. and ZHENG, W. (2008). Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings* **2008** 394–398.

LI, F. (2008). Annual report readability, current earnings, and earnings persistence. *J. Account. Econ.* **45** 221–247. Economic Consequences of Alternative Accounting Standards and Regulation. https://doi.org/10.1016/j.jacceco.2008.02.003

MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J. and MCCLOSKY, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics* (*ACL*) *System Demonstrations*. 55–60.

MARTINC, M., POLLAK, S. and ROBNIK-ŠIKONJA, M. (2021). Supervised and unsupervised neural approaches to text readability. *Comput. Linguist.* **47** 141–179. https://doi.org/10.1162/coli_a_00398

MCCULLAGH, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42** 109–127. MR0583347

MCLAUGHLIN, G. H. (1969). SMOG grading: A new readability formula. *J. Read. Behav.* **12** 639–646.

MURDOCH, W. J., SINGH, C., KUMBIER, K., ABBASI-ASL, R. and YU, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **116** 22071–22080. MR4030584 https://doi.org/10.1073/pnas.1900654116

PARK, M. Y., HASTIE, T. J. and TIBSHIRANI, R. (2007). Averaged gene expressions for regression. *Biostatistics* **8** 212–227. https://doi.org/10.1093/biostatistics/kxl002

PITLER, E. and NENKOVA, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 186–195. Association for Computational Linguistics.

RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1** 206–215. https://doi.org/10.1038/s42256-019-0048-x

SANTOS, R., PEDRO, G. LEAL, S., VALE, O., PARDO, T., BONTCHEVA, K. and SCARTON, C. (2020). Measuring the impact of readability features in fake news detection. In *Proceedings of the* 12*th Language Resources and Evaluation Conference* 1404–1413. European Language Resources Association, Marseille, France.

SCHMIDT, D. (2019). *sylcount*: Syllable counting and readability measurements R package version 0.2-1.

SCHWARM, S. E. and OSTENDORF, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the* 43*rd Annual Meeting on Association for Computational Linguistics*. *ACL* '05 523–530. Association for Computational Linguistics, Stroudsburg, PA, USA. https://doi.org/10.3115/1219840.1219905

SCHWENDINGER, F. and HORNIK, K. (2019). *NLPclient*: Stanford CoreNLP annotation client R package version 1.0.

SCHWENDINGER, F., VANA, L. and HORNIK, K. (2024). Supplement to "Readability prediction: How many features are necessary?" https://doi.org/10.1214/23-AOAS1820SUPPA, https://doi.org/10.1214/23-AOAS1820SUPPB

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250

STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. and HOTHORN, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **8**.

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008). Conditional variable importance for random forests. *BMC Bioinform.* **9** 307. https://doi.org/10.1186/1471-2105-9-307

SUNG, Y.-T., CHEN, J.-L., CHA, J.-H., TSENG, H.-C., CHANG, T.-H. and CHANG, K.-E. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behav. Res. Methods* **47** 340–354. https://doi.org/10.3758/s13428-014-0459-x

TABACHNICK, B. G., FIDELL, L. S. and ULLMAN, J. B. (2007). *Using Multivariate Statistics*, 5th ed. Pearson Education, Boston, MA.

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

VAJJALA, S. and LUČIĆ, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* 297–304. Association for Computational Linguistics, New Orleans, LA. https://doi.org/10.18653/v1/w18-0535

VAJJALA, S. and LUČIĆ, I. (2019). On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* 349–359. https://doi.org/10.18653/v1/w19-4437

VOGEL, M. and WASHBURNE, C. W. (1928). An objective method of determining grade placement of children's reading material. *Elem. Sch. J.* **28** 373–381.

WASHBURNE, C. W. and VOGEL, M. (1926). *Winnetka Graded Book List*. American Library Association, Chicago, IL.

WORRALL, A. P., CONNOLLY, M. J., O'NEILL, A., O'DOHERTY, M., THORNTON, K. P., MCNALLY, C., MCCONKEY, S. J. and DE BARRA, E. (2020). Readability of online Covid-19 health information: A comparison between four English speaking countries. *BMC Public Health* **20** 1–12. https://doi.org/10.1186/s12889-020-09710-5

WRIGHT, M. N. and ZIEGLER, A. (2017). *ranger*: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** 1–17. https://doi.org/10.18637/jss.v077.i01

WURM, M. J., RATHOUZ, P. J. and HANLON, B. M. (2021). Regularized ordinal regression and the ordinalNet R package. *J. Stat. Softw.* **99** 1–42. https://doi.org/10.18637/jss.v099.i06

XU, X. and GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal.* **10** 909–936. MR3432244 https://doi.org/10.1214/14-BA929

YANG, M., REN, Y. and ADOMAVICIUS, G. (2019). Understanding user-generated content and customer engagement on Facebook business pages. *Inf. Syst. Res.* **30** 839–855. https://doi.org/10.1287/isre.2019.0834

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

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1035–1050
https://doi.org/10.1214/23-AOAS1821

INFORMATION-INCORPORATED CLUSTERING ANALYSIS OF DISEASE PREVALENCE TRENDS

BY CHENJIN MA[1,a], CUNJIE LIN[2,b], YUAN XUE[3,c], SANGUO ZHANG[3,d], QINGZHAO ZHANG[4,e] AND SHUANGGE MA[5,f]

[1]*Department of Statistics and Data Science, Beijing University of Technology,* [a]*machenjin@bjut.edu.cn*
[2]*School of Statistics, Renmin University of China,* [b]*lincunjie@ruc.edu.cn*
[3]*School of Mathematics Sciences, University of Chinese Academy of Sciences,* [c]*xueyuan115@mails.ucas.ac.cn,* [d]*sgzhang@ucas.ac.cn*
[4]*Department of Statistics, School of Economics, Xiamen University,* [e]*qzzhang@xmu.edu.cn*
[5]*Department of Biostatistics, Yale School of Public Health,* [f]*shuangge.ma@yale.edu*

In biomedical research the analysis of disease prevalence is of critical importance. While most of the existing prevalence studies focus on individual diseases, there has been increasing effort that jointly examines the prevalence values and their trends of multiple diseases. Such joint analysis can provide valuable insights not shared by individual-disease analysis. A critical limitation of the existing analysis is that there is a lack of attention to existing information, which has been accumulated through a large number of studies and can be valuable especially when there are a large number of diseases but the number of prevalence values for a specific disease is limited. In this study we conduct the functional clustering analysis of prevalence trends for a large number of diseases. A novel approach based on the penalized fusion technique is developed to incorporate information mined from published articles. It is innovatively designed to take into account that such information may not be fully relevant or correct. Another significant development is that statistical properties are rigorously established. Simulation is conducted and demonstrates its competitive performance. In the analysis of data from Taiwan NHIRD (National Health Insurance Research Database), new and interesting findings that differ from the existing ones are made.

REFERENCES

BECKER, K. G., HOSACK, D. A., DENNIS, G., LEMPICKI, R. A., BRIGHT, T. J., CHEADLE, C. and ENGEL, J. (2003). PubMatrix: A tool for multiplex literature mining. *BMC Bioinform.* **4** 1–6.

BIAU, D. J., BOULEZAZ, S., CASABIANCA, L., HAMADOUCHE, M., ANRACT, P. and CHEVRET, S. (2017). Using Bayesian statistics to estimate the likelihood a new trial will demonstrate the efficacy of a new treatment. *BMC Med. Res. Methodol.* **17** 1–10.

CANTO, J. G., SHLIPAK, M. G., ROGERS, W. J., MALMGREN, J. A., FREDERICK, P. D., LAMBREW, C. T., ORNATO, J. P., BARRON, H. V. and KIEFE, C. I. (2000). Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. *JAMA* **283** 3223–3229. https://doi.org/10.1001/jama.283.24.3223

CHEN, L. L., BLUMM, N., CHRISTAKIS, N. A., BARABÁSI, A.-L. and DEISBOECK, T. S. (2009). Cancer metastasis networks and the prediction of progression patterns. *Br. J. Cancer* **101** 749–758. https://doi.org/10.1038/sj.bjc.6605214

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat.* **10** 596–617. MR3528353 https://doi.org/10.1214/16-AOAS912

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory* **57** 5467–5484. MR2849368 https://doi.org/10.1109/TIT.2011.2158486

GABILLOT-CARRÉ, M. and ROUJEAU, J.-C. (2007). Acute bacterial skin infections and cellulitis. *Curr. Opin. Infect. Dis.* **20** 118–123. https://doi.org/10.1097/QCO.0b013e32805dfb2d

Key words and phrases. Clustering, disease prevalence trends, information incorporation, penalized fusion.

GOH, K.-I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. and BARABÁSI, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690.
HSIEH, C.-Y., SU, C.-C., SHAO, S.-C., SUNG, S.-F., LIN, S.-J., YANG, Y.-H. K. and LAI, E. C.-C. (2019). Taiwan's national health insurance research database: Past and future. *Clin. Epidemiol.* **11** 349.
ISCI, S., DOGAN, H., OZTURK, C. and OTU, H. H. (2014). Bayesian network prior: Network analysis of biological data using external knowledge. *Bioinformatics* **30** 860–867. https://doi.org/10.1093/bioinformatics/btt643
JACQUES, J. and PREDA, C. (2014). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y
JADHAV, S., MA, C., JIANG, Y., SHIA, B.-C. and MA, S. (2021). Pan-disease clustering analysis of the trend of period prevalence. *Ann. Appl. Stat.* **15** 1945–1958. MR4355083 https://doi.org/10.1214/21-aoas1470
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
JOFFRES, M., FALASCHETTI, E., GILLESPIE, C., ROBITAILLE, C., LOUSTALOT, F., POULTER, N., MCALISTER, F. A., JOHANSEN, H., BACLIC, O. et al. (2013). Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: A cross-sectional study. *BMJ Open* **3** e003423. https://doi.org/10.1136/bmjopen-2013-003423
KRESSEL, B. R., RYAN, K. P., DUONG, A. T., BERENBERG, J. and SCHEIN, P. S. (1981). Microangiopathic hemolytic anemia, thrombocytopenia, and renal failure in patients treated for adenocarcinoma. *Cancer* **48** 1738–1745. https://doi.org/10.1002/1097-0142(19811015)48:8<1738::aid-cncr2820480808>3.0.co;2-e
LAI, Y.-H. (2015). Network analysis of comorbidities: Case study of HIV/AIDS in Taiwan. In *International Conference on Multidisciplinary Social Networks Research* 174–186. Springer, Berlin.
LUCHSINGER, J. A., REITZ, C., PATEL, B., TANG, M.-X., MANLY, J. J. and MAYEUX, R. (2007). Relation of diabetes to mild cognitive impairment. *Arch. Neurol.* **64** 570–575.
MA, C., LI, Y., SHIA, B. and MA, S. (2020). Human disease cost network analysis. *Stat. Med.* **39** 1237–1249. MR4098487 https://doi.org/10.1002/sim.8472
MA, C., LIN, C., XUE, Y., ZHANG, S., ZHANG, Q. and MA, S. (2024). Supplement to "Information-incorporated clustering analysis of disease prevalence trends." https://doi.org/10.1214/23-AOAS1821SUPPA, https://doi.org/10.1214/23-AOAS1821SUPPB
MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039
MALKIN, D., JOLLY, K. W., BARBIER, N., LOOK, A. T., FRIEND, S. H., GEBHARDT, M. C., ANDERSEN, T. I., BØRRESEN, A.-L., LI, F. P. et al. (1992). Germline mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms. *N. Engl. J. Med.* **326** 1309–1315.
MODESTIN, J., HERMANN, S. and ENDRASS, J. (2007). Schizoidia in schizophrenia spectrum and personality disorders: Role of dissociation. *Psychiatry Res.* **153** 111–118. https://doi.org/10.1016/j.psychres.2006.03.003
RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 305–332. MR2188987 https://doi.org/10.1111/j.1467-9868.2006.00545.x
ROMANOWSKI, M. D., PAROLIN, M. B., FREITAS, A. C. T., PIAZZA, M. J., BASSO, J. and URBANETZ, A. A. (2015). Prevalence of non-alcoholic fatty liver disease in women with polycystic ovary syndrome and its correlation with metabolic syndrome. *Arq. Gastroenterol.* **52** 117–123. https://doi.org/10.1590/S0004-28032015000200008
SCHUMAKER, L. L. (2007). *Spline Functions*: *Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR2348176 https://doi.org/10.1017/CBO9780511618994
SIPPONEN, P. and HYVÄRINEN, H. (1993). Role of Helicobacter pylori in the pathogenesis of gastritis, peptic ulcer and gastric cancer. *Scand. J. Gastroenterol.* **28** 3–6.
TSAI, C.-P., HU, C. and LEE, C. T.-C. (2019). Finding diseases associated with amyotrophic lateral sclerosis: A total population-based case–control study. *Amyotroph. Lateral Scler. Frontotemporal Degeneration* **20** 82–89.
WIEGAND, S., EIVAZI, B., BARTH, P. J., VON RAUTENFELD, D. B., FOLZ, B. J., MANDIC, R. and WERNER, J. A. (2008). Pathogenesis of lymphangiomas. *Virchows Arch.* **453** 1–8. https://doi.org/10.1007/s00428-008-0611-z
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
ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277 https://doi.org/10.1214/aos/1024691356
ZHOU, X., LEI, L., LIU, J., HALU, A., ZHANG, Y., LI, B., GUO, Z., LIU, G., SUN, C. et al. (2018). A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *eBioMedicine* **31** 79–91.

ZHOU, X., MENCHE, J., BARABÁSI, A.-L. and SHARMA, A. (2014). Human symptoms–disease network. *Nat. Commun*. **5** 1–10.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1051–1063
https://doi.org/10.1214/23-AOAS1822

FUNCTIONAL PARTIAL LEAST SQUARES WITH CENSORED OUTCOMES: PREDICTION OF BREAST CANCER RISK WITH MAMMOGRAM IMAGES

BY SHU JIANG[1,a], JIGUO CAO[b] AND GRAHAM A. COLDITZ[c]

[1]*Division of Public Health Sciences, Washington University School of Medicine in St. Louis,* [a]*jiang.shu@wustl.edu,* [b]*jiguo_cao@sfu.ca,* [c]*colditzg@wustl.edu*

We consider the problem of predicting breast cancer risk using mammogram imaging data where the dimension of pixels greatly exceed the number of individuals in the cohort. The functional partial least squares (FPLS) is a popular dimensional reduction method in constructing latent explanatory components using linear combinations of the original predictor variables. While FPLS with scalar responses has been studied in the literature, the presence of right censoring under the survival framework poses challenges in modeling and estimation. Given several different representations for PLS with Cox regression in the literature, we unify and extend three formulations to deal with right censoring, that is, reweighing, mean imputation, and deviance residuals to the functional setting in this paper. We empirically investigate and compare the performance of the three proposed FPLS frameworks in the context of imaging predictor via intensive simulation studies. The proposed methods are applied to the Joanne Knight Breast Health Cohort where we show increased model discriminatory performance under the FPLS framework compared to competing models.

REFERENCES

ANANDARAJAH, A., CHEN, Y., COLDITZ, G. A., HARDI, A., STOLL, C. and JIANG, S. (2022). Studies of parenchymal texture added to mammographic breast density and risk of breast cancer: A systematic review of the methods used in the literature. *Breast Cancer Res*. **24** 1–18.

ANANDARAJAH, A., CHEN, Y., STOLL, C., HARDI, A., JIANG, S. and COLDITZ, G. A. (2023). Repeated measures of mammographic density and texture to evaluate prediction and risk of breast cancer: A systematic review of the methods used in the literature. *Cancer Causes Control* 1–10.

BASTIEN, P., BERTRAND, F., MEYER, N. and MAUMY-BERTRAND, M. (2015). Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics* **31** 397–404.

BOYD, N. F., MARTIN, L. J., ROMMENS, J. M., PATERSON, A. D., MINKIN, S., YAFFE, M. J., STONE, J. and HOPPER, J. L. (2009). Mammographic density: A heritable risk factor for breast cancer. In *Cancer Epidemiology* 343–360.

BRENTNALL, A. R., HARKNESS, E. F., ASTLEY, S. M., DONNELLY, L. S., STAVRINOS, P., SAMPSON, S., FOX, L., SERGEANT, J. C., HARVIE, M. N. et al. (2015). Mammographic density adds accuracy to both the Tyrer–Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res*. **17** 1–10.

CHEN, S., TAMIMI, R. M., COLDITZ, G. A. and JIANG, S. (2023). Association and prediction utilizing craniocaudal and mediolateral oblique view digital mammography and long-term breast cancer risk. *Cancer Prev. Res*. OF1–OF8.

COLDITZ, G. A., BENNETT, D. L., TAPPENDEN, J., BEERS, C., ACKERMANN, N., WU, N., LUO, J., HUMBLE, S., LINNENBRINGER, E. et al. (2022). Joanne Knight breast health cohort at Siteman Cancer Center. *Cancer Causes Control* **33** 623–629.

DATTA, S. (2005). Estimating the mean life time using right censored data. *Stat. Methodol*. **2** 65–69.

DATTA, S., LE-RADEMACHER, J. and DATTA, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63** 259–271.

DEMLER, O. V., PENCINA, M. J. and D'AGOSTINO, R. B. SR. (2012). Misuse of DeLong test to compare AUCs for nested models. *Stat. Med*. **31** 2577–2587.

Key words and phrases. Functional partial least squares, survival analysis, image analysis, risk prediction.

GAIL, M. H., BRINTON, L. A., BYAR, D. P., CORLE, D. K., GREEN, S. B., SCHAIRER, C. and MULVIHILL, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81** 1879–1886. https://doi.org/10.1093/jnci/81.24.1879

GERDS, T. A., CAI, T. and SCHUMACHER, M. (2008). The performance of risk prediction models. *Biom. J.* **50** 457–479.

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/10.1002/1097-0258

JIANG, S., CAO, J. and COLDITZ, G. A (2024). Supplement to "Functional partial least squares with censored outcomes: Prediction of breast cancer risk with mammogram images." https://doi.org/10.1214/23-AOAS1822SUPP

JIANG, S., CAO, J., COLDITZ, G. A. and ROSNER, B. (2023a). Predicting the onset of breast cancer using mammogram imaging data with irregular boundary. *Biostatistics* **24** 358–371.

JIANG, S., CAO, J., ROSNER, B. and COLDITZ, G. A. (2023b). Supervised two-dimensional functional principal component analysis with time-to-event outcomes and mammogram imaging data. *Biometrics* **79** 1359–1369.

JIANG, S. and COLDITZ, G. A. (2023). Causal mediation analysis using high-dimensional image mediator bounded in irregular domain with an application to breast cancer. *Biometrics*. https://doi.org/10.1111/biom.13847

KESHARI, R., VATSA, M., SINGH, R. and NOORE, A. (2018). Learning structure and strength of CNN filters for small sample size training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 9349–9358.

KONG, D., IBRAHIM, J. G., LEE, E. and ZHU, H. (2018). FLCRM: Functional linear Cox regression model. *Biometrics* **74** 109–117.

LEDELL, E., PETERSEN, M. and VAN DER LAAN, M. (2015). Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9** 1583.

LI, H. and GUI, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* **20** i208–i215.

MAAS, P., BARRDAHL, M., JOSHI, A. D., AUER, P. L., GAUDET, M. M., MILNE, R. L., SCHUMACHER, F. R., ANDERSON, W. F., CHECK, D. et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2** 1295–1302.

MARTENS, H. and NÆS, T. (1992). *Multivariate Calibration*. Wiley, Chichester.

NYGÅRD, S., BORGAN, Ø., LINGJÆRDE, O. C. and STØRVOLD, H. L. (2008). Partial least squares Cox regression for genome-wide data. *Lifetime Data Anal.* **14** 179–195.

PARK, P. J., TIAN, L. and KOHANE, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* **18** S120–S127.

PASHAYAN, N., MORRIS, S., GILBERT, F. J. and PHAROAH, P. D. P. (2018). Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: A life-table model. *JAMA Oncol.* **4** 1504–1510.

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

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

ROSNER, B., TAMIMI, R. M., KRAFT, P., GAO, C., MU, Y., SCOTT, C., WINHAM, S. J., VACHON, C. M. and COLDITZ, G. A. (2021). Simplified breast risk tool integrating questionnaire risk factors, mammographic density, and polygenic risk score: Development and validation. *Cancer Epidemiol. Biomark. Prev.* **30** 600–607.

SEGAL, M. R. (2006). Microarray gene expression data with linked survival phenotypes: Diffuse large-B-cell lymphoma revisited. *Biostatistics* **7** 268–285.

TABAR, L., GAD, A., HOLMBERG, L., LJUNGQUIST, U., FAGERBERG, C., BALDETORP, L., GRÖNTOFT, O., LUNDSTRÖM, B., MÅNSON, J. et al. (1985). Reduction in mortality from breast cancer after mass screening with mammography: Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* **325** 829–832.

UNO, H., CAI, T., TIAN, L. and WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* **102** 527–537.

VICKERS, A. J., CRONIN, A. M. and BEGG, C. B. (2011). One statistical test is sufficient for assessing new predictive markers. *BMC Med. Res. Methodol.* **11** 1–7.

VILMUN, B. M., VEJBORG, I., LYNGE, E., LILLHOLM, M., NIELSEN, M., NIELSEN, M. B. and CARLSEN, J. F. (2020). Impact of adding breast density to breast cancer risk models: A systematic review. *Eur. J. Radiol.* **127** 109019.

VISVANATHAN, K., FABIAN, C. J., BANTUG, E., BREWSTER, A. M., DAVIDSON, N. E., DECENSI, A., FLOYD, J. D., GARBER, J. E., HOFSTATTER, E. W. et al. (2019). Use of endocrine therapy for breast cancer risk reduction: ASCO clinical practice guideline update. *J. Clin. Oncol.* **37** 3152–3165.

WAGNER, R., THOM, M., SCHWEIGER, R., PALM, G. and ROTHERMEL, A. (2013). Learning convolutional neural networks from few samples. In *The* 2013 *International Joint Conference on Neural Networks* (*IJCNN*) 1–7. IEEE Press, New York.

WOLD, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis* (*Proc*. *Internat*. *Sympos*., *Dayton*, *Ohio*, 1965) 391–420. Academic Press, New York.

WOLD, H. (1975a). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. *J*. *Appl*. *Probab*. **12** 117–142.

WOLD, H. (1975b). Path models with latent variables: The NIPALS approach. In *Quantitative Sociology* 307–357.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1064–1089
https://doi.org/10.1214/23-AOAS1823

EFFICIENT AND EFFECTIVE CALIBRATION OF NUMERICAL MODEL OUTPUTS USING HIERARCHICAL DYNAMIC MODELS

BY YEWEN CHEN[1,a], XIAOHUI CHANG[2,b], BOHAI ZHANG[3,c] AND HUI HUANG[4,d]

[1]*College of Public Health, University of Georgia,* [a]*Yewen.Chen@uga.edu*

[2]*College of Business, Oregon State University,* [b]*xiaohui.chang@oregonstate.edu*

[3]*Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College,* [c]*bohaizhang@uic.edu.cn*

[4]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [d]*huangh89@mail.sysu.edu.cn*

Numerical air quality models, such as the Community Multiscale Air Quality (CMAQ) system, play a critical role in characterizing pollution levels at fine spatial and temporal scales. The model outputs, however, tend to systematically over- or underestimate the real pollutant concentrations. In this study we propose a Bayesian hierarchical dynamic model to calibrate large-scale grid-level CMAQ model outputs using data from other sources, especially point-level observations from sparsely located monitoring stations. In our model a stochastic integro-differential equation (IDE) is implemented to account for space-time interactions of air pollutants. To better approximate the spatial pattern of pollutants, we employ nonregular meshes to discretize IDEs. A spatial partitioning procedure is embedded to improve the scalability of the approach for very large meshes. An algorithm based on variational Bayes and ensemble Kalman smoother is developed to accelerate the parameter estimation and calibration procedure. We apply the proposed approach to calibrate CMAQ outputs for China's Beijing–Tianjin–Hebei region. In contrast to existing methods, the proposed approach captures space-time interactions, produces more accurate calibration results, and operates at a higher computational efficiency. A reanalysis dataset is also adopted to demonstrate the effectiveness and efficiency of our approach to large spatial data.

REFERENCES

APPEL, K. W., NAPELENOK, S. L., FOLEY, K. M., PYE, H. O. T., HOGREFE, C., LUECKEN, D. J., BASH, J. O., ROSELLE, S. J., PLEIM, J. E. et al. (2017). Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geosci. Model Dev.* **10** 1703–1732.

BAKAR, K. S., KOKIC, P. and JIN, H. (2016). Hierarchical spatially varying coefficient and temporal dynamic process models using spTDyn. *J. Stat. Comput. Simul.* **86** 820–840. MR3432519 https://doi.org/10.1080/00949655.2015.1038267

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 825–848. MR2523906 https://doi.org/10.1111/j.1467-9868.2008.00663.x

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. MR2787270 https://doi.org/10.1007/s13253-009-0004-z

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2012). Space-time data fusion under error in computer model output: An application to modeling air quality. *Biometrics* **68** 837–848. MR3055188 https://doi.org/10.1111/j.1541-0420.2011.01725.x

BERROCAL, V. J., GUAN, Y., MUYSKENS, A., WANG, H., REICH, B. J., MULHOLLAND, J. A. and CHANG, H. H. (2020). A comparison of statistical and machine learning methods for creating national daily maps of ambient $PM_{2.5}$ concentration. *Atmos. Environ.* **222** 117130.

Key words and phrases. Calibration, numerical model outputs, hierarchical dynamic models, stochastic integro-differential equations, variational Bayes, space-partitioning-based ensemble Kalman smoother.

BLANGIARDO, M. and CAMELETTI, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. Wiley, Chichester. MR3364017

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BOLIN, D., WALLIN, J. and LINDGREN, F. (2019). Latent Gaussian random field mixture models. *Comput. Statist. Data Anal.* **130** 80–93. MR3860530 https://doi.org/10.1016/j.csda.2018.08.007

BYUN, D. and SCHERE, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Appl. Mech. Rev.* **59** 51–77.

CHANG, J. C. and HANNA, S. R. (2004). Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **87** 167–196.

CHEN, J. and STEIN, M. L. (2023). Linear-cost covariance functions for Gaussian random fields. *J. Amer. Statist. Assoc.* **118** 147–164. MR4571113 https://doi.org/10.1080/01621459.2021.1919122

CHEN, Y., CHANG, X., LUO, F. and HUANG, H. (2023). Additive dynamic models for correcting numerical model outputs. *Comput. Statist. Data Anal.* **187** Paper No. 107799, 21 pp. MR4604820 https://doi.org/10.1016/j.csda.2023.107799

CHEN, Y., CHANG, X., ZHANG, B. and HUANG, H. (2024). Supplement to "Efficient and effective calibration of numerical model outputs using hierarchical dynamic models." https://doi.org/10.1214/23-AOAS1823SUPPA, https://doi.org/10.1214/23-AOAS1823SUPPB

CHINA'S STATE COUNCIL (2013). The action plan for air pollution prevention and control. Available at http://www.gov.cn/zwgk/2013-09/12/content_2486773.htm. In Chinese.

CHINA'S STATE COUNCIL (2018). The three-year action plan for winning the blue sky defense battle. Available at http://www.gov.cn/xinwen/2018-07/03/content_5303212.htm. In Chinese.

CHINA'S STATE COUNCIL (2021). The long-term "Beautiful China" targets through 2035. Available at http://www.gov.cn/xinwen/2021-02/22/content_5588304.htm. In Chinese.

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

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

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res., Oceans* **99** 10143–10162.

EVENSEN, G. and VAN LEEUWEN, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Mon. Weather Rev.* **128** 1852–1867.

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199 https://doi.org/10.1111/j.0006-341X.2005.030821.x

FURRER, R., GENTON, M. G. and NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15** 502–523. MR2291261 https://doi.org/10.1198/106186006X132178

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

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

GUAN, Y., JOHNSON, M. C., KATZFUSS, M., MANNSHARDT, E., MESSIER, K. P., REICH, B. J. and SONG, J. J. (2020). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. *J. Amer. Statist. Assoc.* **115** 1111–1124. MR4143453 https://doi.org/10.1080/01621459.2019.1665526

GUILLAS, S., BAO, J., CHOI, Y. and WANG, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmos. Environ.* **42** 1338–1348.

HAN, L., CHEN, M., CHEN, K., CHEN, H., ZHANG, Y., LU, B., SONG, L. and QIN, R. (2021). A deep learning method for bias correction of ECMWF 24–240 h forecasts. *Adv. Atmos. Sci.* **38** 1444–1459.

HARVEY, D., LEYBOURNE, S. and NEWBOLD, P. (1997). Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **13** 281–291.

HE, D., ZHOU, Z., KANG, Z. and LIU, L. (2019). Numerical studies on forecast error correction of GRAPES model with variational approach. *Adv. Meteorol.* **2019** 1–13.

HEATON, M. J., CHRISTENSEN, W. F. and TERRES, M. A. (2017). Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics* **59** 93–101. MR3604192 https://doi.org/10.1080/00401706.2015.1102763

HEATON, M. J., DATTA, A., FINLEY, A. O. et al. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* **24** 398–425. MR3996451 https://doi.org/10.1007/s13253-018-00348-w

HEINRICH, C., HELLTON, K. H., LENKOSKI, A. and THORARINSDOTTIR, T. L. (2021). Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *J. Amer. Statist. Assoc.* **116** 1048–1059. MR4309249 https://doi.org/10.1080/01621459.2020.1769634

HERSBACH, H., BELL, B., BERRISFORD, P., HIRAHARA, S., HORÁNYI, A., MUÑOZ-SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R. et al. (2020). The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146** 1999–2049.

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

HOUTEKAMER, P. L., MITCHELL, H. L., PELLERIN, G., BUEHNER, M., CHARRON, M., SPACEK, L. and HANSEN, B. (2005). Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Weather Rev.* **133** 604–620.

HOUTEKAMER, P. L. and ZHANG, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **144** 4489–4532.

ISHIGURO, K., SATO, I. and UEDA, N. (2017). Averaged collapsed variational Bayes inference. *J. Mach. Learn. Res.* **18** Paper. No. 1, 29 pp. MR3625705

JIANG, X. and YOO, E.-H. E. (2019). Modeling wildland fire-specific $PM_{2.5}$ concentrations for uncertainty-aware health impact assessments. *Environ. Sci. Technol.* **53** 11828–11839. https://doi.org/10.1021/acs.est.9b02660

KATZFUSS, M., STROUD, J. R. and WIKLE, C. K. (2016). Understanding the ensemble Kalman filter. *Amer. Statist.* **70** 350–357. MR3574787 https://doi.org/10.1080/00031305.2016.1141709

KATZFUSS, M., STROUD, J. R. and WIKLE, C. K. (2020). Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models. *J. Amer. Statist. Assoc.* **115** 866–885. MR4107685 https://doi.org/10.1080/01621459.2019.1592753

KAUFMAN, C. G., SCHERVISH, M. J. and NYCHKA, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *J. Amer. Statist. Assoc.* **103** 1545–1555. MR2504203 https://doi.org/10.1198/016214508000000959

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

KIM, H.-M., MALLICK, B. K. and HOLMES, C. C. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *J. Amer. Statist. Assoc.* **100** 653–668. MR2160567 https://doi.org/10.1198/016214504000002014

KIRCHGESSNER, P., NERGER, L. and BUNSE-GERSTNER, A. (2014). On the choice of an optimal localization radius in ensemble Kalman filter methods. *Mon. Weather Rev.* **142** 2165–2175.

KONG, L., TANG, X., ZHU, J., WANG, Z., WU, H. and LI, J. (2020). Developing high-resolution air quality reanalysis dataset over China for years 2013–2018 based on ensemble Kalman filter and surface observations from CNEMC. In *EGU General Assembly Conference Abstracts* 6848.

KONOMI, B. A., SANG, H. and MALLICK, B. K. (2014). Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *J. Comput. Graph. Statist.* **23** 802–829. MR3224657 https://doi.org/10.1080/10618600.2013.812872

KOT, M., LEWIS, M. A. and VAN DEN DRIESSCHE, P. (1996). Dispersal data and the spread of invading organisms. *Ecology* **77** 2027–2042.

LIANG, D., ZHANG, H., CHANG, X. and HUANG, H. (2021). Modeling and regionalization of China's $PM_{2.5}$ using spatial-functional mixture models. *J. Amer. Statist. Assoc.* **116** 116–132. MR4227679 https://doi.org/10.1080/01621459.2020.1764363

LIANG, X., ZOU, T., GUO, B., LI, S., ZHANG, H., ZHANG, S., HUANG, H. and CHEN, S. X. (2015). Assessing Beijing's $PM_{2.5}$ pollution: Severity, weather impact, APEC and winter heating. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **471** 20150257.

LINDGREN, F. and RUE, H. (2015). Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63** 1–25.

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

LU, X., ZHANG, S., XING, J., WANG, Y., CHEN, W., DING, D., WU, Y., WANG, S., DUAN, L. et al. (2020). Progress of air pollution control in China and its challenges and opportunities in the ecological civilization era. *Engineering* **6** 1423–1431.

MCMILLAN, N. J., HOLLAND, D. M., MORARA, M. and FENG, J. (2010). Combining numerical model output and particulate data using Bayesian space-time modeling. *Environmetrics* **21** 48–65. MR2842223 https://doi.org/10.1002/env.984

MITCHELL, H. L., HOUTEKAMER, P. L. and PELLERIN, G. (2002). Ensemble size, balance, and model-error representation in an ensemble Kalman filter. *Mon. Weather Rev.* **130** 2791–2808.

NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Statist.* **24** 579–599. MR3357396 https://doi.org/10.1080/10618600.2014.914946

QI, J., ZHENG, B., LI, M., YU, F., CHEN, C., LIU, F., ZHOU, X., YUAN, J., ZHANG, Q. et al. (2017). A high-resolution air pollutants emission inventory in 2013 for the Beijing–Tianjin–Hebei region, China. *Atmos. Environ.* **170** 156–168.

REN, Q., BANERJEE, S., FINLEY, A. O. and HODGES, J. S. (2011). Variational Bayesian methods for spatial data analysis. *Comput. Statist. Data Anal.* **55** 3197–3217. MR2825404 https://doi.org/10.1016/j.csda.2011.05.021

RICHARDSON, R., KOTTAS, A. and SANSÓ, B. (2017). Flexible integro-difference equation modeling for spatio-temporal data. *Comput. Statist. Data Anal.* **109** 182–198. MR3603648 https://doi.org/10.1016/j.csda.2016.11.011

RODU, J. and KAFADAR, K. (2022). The q-q boxplot. *J. Comput. Graph. Statist.* **31** 26–39. MR4387208 https://doi.org/10.1080/10618600.2021.1938586

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

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

SAHU, S. K., GELFAND, A. E. and HOLLAND, D. M. (2006). Spatiotemporal modeling of fine particulate matter. *J. Agric. Biol. Environ. Stat.* **11** 61–86.

SALTER, J. M., WILLIAMSON, D. B., SCINOCCA, J. and KHARIN, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *J. Amer. Statist. Assoc.* **114** 1800–1814. MR4047301 https://doi.org/10.1080/01621459.2018.1514306

SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. MR2907125 https://doi.org/10.1214/11-AOAS478

SHADDICK, G., THOMAS, M. L., GREEN, A. et al. (2018). Data integration model for air quality: A hierarchical approach to the global estimation of exposures to ambient air pollution. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 231–253. MR3758764 https://doi.org/10.1111/rssc.12227

STROUD, J. R., STEIN, M. L., LESHT, B. M., SCHWAB, D. J. and BELETSKY, D. (2010). An ensemble Kalman filter and smoother for satellite data assimilation. *J. Amer. Statist. Assoc.* **105** 978–990. MR2752594 https://doi.org/10.1198/jasa.2010.ap07636

TABOUY, T., BARBILLON, P. and CHIQUET, J. (2020). Variational inference for stochastic block models from sampled data. *J. Amer. Statist. Assoc.* **115** 455–466. MR4078475 https://doi.org/10.1080/01621459.2018.1562934

VANNITSEM, S., BREMNES, J. B., DEMAEYER, J., EVANS, G. R., FLOWERDEW, J., HEMRI, S., LERCH, S., ROBERTS, N., THEIS, S. et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Am. Meteorol. Soc.* **102** E681–E699.

WAN, Y., XU, M., HUANG, H. and CHEN, S. X. (2021). A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. *Environmetrics* **32** Paper No. e2648, 16 pp. MR4207558 https://doi.org/10.1002/env.2648

WANG, Y., DU, Y., WANG, J. and LI, T. (2019). Calibration of a low-cost $PM_{2.5}$ monitor using a random forest model. *Environ. Int.* **133** 105161.

WANG, Z. F., XIE, F. Y., WANG, X. Q., AN, J. and ZHU, J. (2006). Development and application of nested air quality prediction modeling system (in Chinese). *Chin. J. Atmos. Sci.* **30** 778–790.

WENDLAND, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.* **4** 389–396. MR1366510 https://doi.org/10.1007/BF02123482

WIKLE, C. K. (2002). A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Stat. Model.* **2** 299–314. MR1951587 https://doi.org/10.1191/1471082x02st036oa

WIKLE, C. K. and HOLAN, S. H. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *J. Time Series Anal.* **32** 339–350. MR2841788 https://doi.org/10.1111/j.1467-9892.2011.00729.x

WIKLE, C. K., ZAMMIT-MANGION, A. and CRESSIE, N. (2019). *Spatiotemporal Statistics with R*. CRC Press/CRC, Boca Raton.

WILKS, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Academic Press, Waltham, MA.

XU, K., WIKLE, C. K. and FOX, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *J. Amer. Statist. Assoc.* **100** 1133–1144. MR2236929 https://doi.org/10.1198/016214505000000682

ZAMMIT-MANGION, A. and WIKLE, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spat. Stat.* **37** 100408, 20 pp. MR4109596 https://doi.org/10.1016/j.spasta.2020.100408

ZHANG, B., SANG, H. and HUANG, J. Z. (2015). Full-scale approximations of spatio-temporal covariance models for large datasets. *Statist. Sinica* **25** 99–114. MR3328805

ZHANG, L., SHAO, J., LU, X., ZHAO, Y., HU, Y., HENZE, D. K., LIAO, H., GONG, S. and ZHANG, Q. (2016). Sources and processes affecting fine particulate matter pollution over North China: An adjoint analysis of the Beijing APEC period. *Environ. Sci. Technol.* **50** 8731–8740.

ZHAO, C., WANG, Q., BAN, J., LIU, Z., ZHANG, Y., MA, R., LI, S. and LI, T. (2020). Estimating the daily $PM_{2.5}$ concentration in the Beijing–Tianjin–Hebei region using a random forest model with a $0.01^\circ \times 0.01^\circ$ spatial resolution. *Environ. Int.* **134** 105297.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1090–1112
https://doi.org/10.1214/23-AOAS1824

NETWORK METHOD FOR VOXEL-PAIR-LEVEL BRAIN CONNECTIVITY ANALYSIS UNDER SPATIAL-CONTIGUITY CONSTRAINTS

BY TONG LU[1,a], YUAN ZHANG[2,b], PETER KOCHUNOV[3,c], ELLIOT HONG[3,d] AND SHUO CHEN[3,e]

[1]*Department of Mathematics, University of Maryland,* [a]*hitonggg@terpmail.umd.edu*
[2]*Department of Statistics, The Ohio State University,* [b]*yzhanghf@stat.osu.edu*
[3]*Maryland Psychiatric Research Center, School of Medicine, University of Maryland,* [c]*pkochunov@som.umaryland.edu,* [d]*pkhong@som.umaryland.edu,* [e]*shuochen@som.umaryland.edu*

Brain connectome analysis commonly compresses high-resolution brain scans (typically composed of millions of voxels) down to only hundreds of *regions of interest* (ROIs) by averaging within-ROI signals. This significant dimension reduction improves computational speed and the morphological properties of anatomical structures; however, it comes at the cost of substantial losses in spatial specificity and sensitivity, especially when the signals exhibit high within-ROI heterogeneity. Oftentimes, abnormally expressed *functional connectivity* (FC) between a pair of ROIs, caused by a brain disease, is primarily driven by only small subsets of voxel pairs within the ROI pair. This article proposes a new network method for the detection of voxel-pair-level neural dysconnectivity with spatial constraints. Specifically, focusing on an ROI pair, our model aims to extract dense subareas that contain aberrant voxel-pair connections while ensuring that the involved voxels are spatially contiguous. In addition, we develop subcommunity-detection algorithms to realize the model, and we justify the consistency of these algorithms. Comprehensive simulation studies demonstrate our method's effectiveness in reducing the false-positive rate while increasing statistical power, detection replicability, and spatial specificity. We apply our approach to reveal: (i) disrupted voxelwise FC patterns related to nicotine addiction between the basal ganglia, hippocampus, and insular gyrus in 3269 participants using UK Biobank data; (ii) voxelwise schizophrenia-altered FC patterns within the salience and temporal-thalamic network in 330 participants in a schizophrenia study. The detected results align with previous medical findings but include improved localized information.

REFERENCES

AGOSTA, F., SALA, S., VALSASINA, P., MEANI, A., CANU, E., MAGNANI, G., CAPPA, S. F., SCOLA, E., QUATTO, P. et al. (2013). Brain network connectivity assessed using graph theory in frontotemporal dementia. *Neurology* **81** 134–143.

ARNOLD, S. E. and TROJANOWSKI, J. Q. (1996). Recent advances in defining the neuropathology of schizophrenia. *Acta Neuropathol.* **92** 217–231. https://doi.org/10.1007/s004010050512

BAHRAMI, M., LAURIENTI, P. J. and SIMPSON, S. L. (2019). Analysis of brain subnetworks within the context of their whole-brain networks. *Hum. Brain Mapp.* **40** 5123–5141. https://doi.org/10.1002/hbm.24762

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

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10** 186–198. https://doi.org/10.1038/nrn2575

CAO, M., WANG, J.-H., DAI, Z.-J., CAO, X.-Y., JIANG, L.-L., FAN, F.-M., SONG, X.-W., XIA, M.-R., SHU, N. et al. (2014). Topological organization of the human brain functional connectome across the lifespan. *Dev. Cogn. Neurosci.* **7** 76–93.

Key words and phrases. Brain connectome, spatial contiguity, voxel-pair-level connectivity, fMRI.

ÇETIN, M. S., CHRISTENSEN, F., ABBOTT, C. C., STEPHEN, J. M., MAYER, A. R., CAÑIVE, J. M., BUSTILLO, J. R., PEARLSON, G. D. and CALHOUN, V. D. (2014). Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia. *NeuroImage* **97** 117–126. https://doi.org/10.1016/j.neuroimage.2014.04.009

CHEN, S., BOWMAN, F. D. and MAYBERG, H. S. (2016). A Bayesian hierarchical framework for modeling brain connectivity for neuroimaging data. *Biometrics* **72** 596–605. MR3515786 https://doi.org/10.1111/biom.12433

CRADDOCK, R. C., JAMES, G. A., HOLTZHEIMER III, P. E., HU, X. P. and MAYBERG, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* **33** 1914–1928.

DERADO, G., BOWMAN, F. D. and KILTS, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics* **66** 949–957. MR2758231 https://doi.org/10.1111/j.1541-0420.2009.01355.x

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

EICKHOFF, S. B., THIRION, B., VAROQUAUX, G. and BZDOK, D. (2015). Connectivity-based parcellation: Critique and implications. *Hum. Brain Mapp.* **36** 4771–4792. https://doi.org/10.1002/hbm.22933

ERSCHE, K. D., BARNES, A., JONES, P. S., MOREIN-ZAMIR, S., ROBBINS, T. W. and BULLMORE, E. T. (2011). Abnormal structure of frontostriatal brain systems is associated with aspects of impulsivity and compulsivity in cocaine dependence. *Brain* **134** 2013–2024.

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

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, L., LI, H., ZHUO, J., ZHANG, Y., WANG, J., CHEN, L., YANG, Z., CHU, C., XIE, S. et al. (2016). The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cereb. Cortex* **26** 3508–3526.

FEDOTA, J. R. and STEIN, E. A. (2015). Resting-state functional connectivity and nicotine addiction: Prospects for biomarker development. *Ann. N.Y. Acad. Sci.* **1349** 64–82. https://doi.org/10.1111/nyas.12882

FERRI, J., FORD, J., ROACH, B., TURNER, J., VAN ERP, T., VOYVODIC, J., PREDA, A., BELGER, A., BUSTILLO, J. et al. (2018). Resting-state thalamic dysconnectivity in schizophrenia and relationships with symptoms. *Psychol. Med.* **48** 2492–2499.

FORNITO, A., ZALESKY, A. and BULLMORE, E. (2016). *Fundamentals of Brain Network Analysis*. Academic Press, San Diego.

GAZNICK, N., TRANEL, D., MCNUTT, A. and BECHARA, A. (2014). Basal ganglia plus insula damage yields stronger disruption of smoking addiction than basal ganglia damage alone. *Nicotine Tob. Res.* **16** 445–453. https://doi.org/10.1093/ntr/ntt172

GRÜNWALD, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge.

GUPTA, J. K., SINGH, S. and VERMA, N. K. (2013). MTBA: MATLAB Toolbox for Biclustering Analysis. 94-97. IEEE.

KAMVAR, S., KLEIN, D. and MANNING, C. (2003). Spectral Learning Technical Report No. 2003-25 Stanford InfoLab.

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

LOEWE, K., GRUESCHOW, M., STOPPEL, C. M., KRUSE, R. and BORGELT, C. (2014). Fast construction of voxel-level functional connectivity graphs. *BMC Neurosci.* **15** 1–13.

LU, T., ZHANG, Y., KOCHUNOV, P., HONG, E. and CHEN, S. (2024). Supplement to "Network method for voxel-pair-level brain connectivity analysis under spatial-contiguity constraints." https://doi.org/10.1214/23-AOAS1824SUPPA, https://doi.org/10.1214/23-AOAS1824SUPPB

LYNALL, M.-E., BASSETT, D. S., KERWIN, R., MCKENNA, P. J., KITZBICHLER, M., MULLER, U. and BULLMORE, E. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* **30** 9477–9487. https://doi.org/10.1523/JNEUROSCI.0333-10.2010

MCCLERNON, F. J., CONKLIN, C. A., KOZINK, R. V., ADCOCK, R. A., SWEITZER, M. M., ADDICOTT, M. A., CHOU, Y.-H., CHEN, N.-K., HALLYBURTON, M. B. et al. (2016). Hippocampal and insular response to smoking-related environments: Neuroimaging evidence for drug-context effects in nicotine dependence. *Neuropsychopharmacology* **41** 877–885.

NICHOLS, T. E. and HOLMES, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15** 1–25. https://doi.org/10.1002/hbm.1058

PALANIYAPPAN, L., WHITE, T. P. and LIDDLE, P. F. (2012). The concept of salience network dysfunction in schizophrenia: From neuroimaging observations to therapeutic opportunities. *Curr. Top. Med. Chem.* **12** 2324–2338. https://doi.org/10.2174/156802612805289881

ROGERS, B. P., MORGAN, V. L., NEWTON, A. T. and GORE, J. C. (2007). Assessing functional connectivity in the human brain by fMRI. *Magn. Reson. Imaging* **25** 1347–1357. https://doi.org/10.1016/j.mri.2007.03.007

RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069. https://doi.org/10.1016/j.neuroimage.2009.10.003

SIMPSON, S. L., BOWMAN, F. D. and LAURIENTI, P. J. (2013). Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Stat. Surv.* **7** 1–36. MR3161730 https://doi.org/10.1214/13-SS103

SUTHERLAND, M. T. and STEIN, E. A. (2018). Functional neurocircuits and neuroimaging biomarkers of tobacco use disorder. *Trends Mol. Med.* **24** 129–143. https://doi.org/10.1016/j.molmed.2017.12.002

THIRION, B., FLANDIN, G., PINEL, P., ROCHE, A., CIUCIU, P. and POLINE, J.-B. (2006). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* **27** 678–693. https://doi.org/10.1002/hbm.20210

WIG, G. S., LAUMANN, T. O. and PETERSEN, S. E. (2014). An approach for parcellating human cortical areas using resting-state correlations. *NeuroImage* **93** 276–291. https://doi.org/10.1016/j.neuroimage.2013.07.035

WU, G.-R., STRAMAGLIA, S., CHEN, H., LIAO, W. and MARINAZZO, D. (2013). Mapping the voxel-wise effective connectome in resting state fMRI. *PLoS ONE* **8** e73670.

WU, T., WANG, L., HALLETT, M., CHEN, Y., LI, K. and CHAN, P. (2011). Effective connectivity of brain networks during self-initiated movement in Parkinson's disease. *NeuroImage* **55** 204–215.

WYLIE, K. P. and TREGELLAS, J. R. (2010). The role of the insula in schizophrenia. *Schizophr. Res.* **123** 93–104. https://doi.org/10.1016/j.schres.2010.08.027

XIA, Y. and LI, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics* **73** 780–791. MR3713112 https://doi.org/10.1111/biom.12633

XIA, Y. and LI, L. (2019). Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Statist. Sinica* **29** 303–328. MR3889369

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1113–1136
https://doi.org/10.1214/23-AOAS1825

A POPULATION-AWARE RETROSPECTIVE REGRESSION TO DETECT GENOME-WIDE VARIANTS WITH SEX DIFFERENCE IN ALLELE FREQUENCY

BY ZHONG WANG[1,a], ANDREW D. PATERSON[2,b] AND LEI SUN[3,c]

[1]*Department of Statistics and Data Science, Faculty of Science, National University of Singapore,* [a]*zhongwang857@gmail.com*
[2]*Genetics and Genome Biology, The Hospital for Sick Children,* [b]*andrew.paterson@sickkids.ca*
[3]*Biostatistics Division, Dalla Lana School of Public Health, University of Toronto,* [c]*lei.sun@utoronto.ca*

Sex difference in allele frequency is an emerging topic that is crucial to our understanding of data quality and features, particularly when it comes to the largely overlooked X chromosome. To detect sex differences in allele frequency for both X chromosomal and autosomal variants, the existing method is conservative when applied to samples from multiple ancestral populations. Additionally, it remains unexplored whether the sex difference in allele frequency varies between populations, which is important for transancestral genetic studies. To answer these questions, we thus developed a novel, retrospective regression-based testing framework that led to interpretable and easy-to-implement solutions. We then applied the proposed methods to the high-coverage whole genome sequence data of the 1000 Genomes Project, robustly analyzing all samples available from the five super-populations. We had 97 novel findings by recognizing and modelling ancestral differences. Finally, we replicated the specific findings and overall conclusion using the gnomAD v3.1.2 data.

REFERENCES

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. and ZONDERVAN, K. T. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc.* **5** 1564–1573. https://doi.org/10.1038/nprot.2010.116

BACANU, S.-A., DEVLIN, B. and ROEDER, K. (2002). Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* **22** 78–93. https://doi.org/10.1002/gepi.1045

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. and ROTHSTEIN, H. R. (2021). *Introduction to Meta-Analysis*. Wiley, New York.

BROWNING, B. L., TIAN, X., ZHOU, Y. and BROWNING, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108** 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

BYRSKA-BISHOP, M., EVANI, U. S., ZHAO, X., BASILE, A. O., ABEL, H. J., REGIER, A. A., CORVELO, A., CLARKE, W. E., MUSUNURI, R. et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185** 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

CHEN, B., CRAIU, R. V., STRUG, L. J. and SUN, L. (2021). The X factor: A robust and powerful approach to X-chromosome-inclusive whole-genome association studies. *Genet. Epidemiol.* **45** 694–709. https://doi.org/10.1002/gepi.22422

CHEN, C.-F. (1983). Score tests for regression models. *J. Amer. Statist. Assoc.* **78** 158–161. https://doi.org/10.1080/01621459.1983.10477945

CHEN, S., FRANCIOLI, L. C., GOODRICH, J. K., COLLINS, R. L., KANAI, M., WANG, Q., ALFÖLDI, J., WATTS, N. A. and VITTAL, C. (2022). A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv 2022–03.

CROW, J. F. and KIMURA, M. (1970). *An Introduction to Population Genetics Theory*. Harper & Row, New York. MR0274068

DAS, S., FORER, L., SCHÖNHERR, S., SIDORE, C., LOCKE, A. E., KWONG, A., VRIEZE, S. I., CHEW, E. Y., LEVY, S. et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* **48** 1284–1287.

Key words and phrases. sdMAF test, regression.

DERKACH, A., LAWLESS, J. F. and SUN, L. (2014). Pooled association tests for rare genetic variants: A review and some new results. *Statist. Sci.* **29** 302–321. MR3264544 https://doi.org/10.1214/13-STS456

DUDBRIDGE, F. and GUSNANTO, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32** 227–234. https://doi.org/10.1002/gepi.20297

KÖNIG, I. R., LOLEY, C., ERDMANN, J. and ZIEGLER, A. (2014). How to include chromosome X in your genome-wide association study. *Genet. Epidemiol.* **38** 97–103. https://doi.org/10.1002/gepi.21782

LIN, D. Y. and ZENG, D. (2009). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genet. Epidemiol.* https://doi.org/10.1002/gepi.20435

LIN, D. Y. and ZENG, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97** 321–332. MR2650741 https://doi.org/10.1093/biomet/asq006

MAREES, A. T., DE KLUIVER, H., STRINGER, S., VORSPAN, F., CURIS, E., MARIE-CLAIRE, C. and DERKS, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27** e1608. https://doi.org/10.1002/mpr.1608

PIRASTU, N., CORDIOLI, M., NANDAKUMAR, P., MIGNOGNA, G., ABDELLAOUI, A., HOLLIS, B., KANAI, M., RAJAGOPAL, V. M. and PAROLO, P. D. B. et al. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* **53** 663–671.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909. https://doi.org/10.1038/ng1847

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. A. et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Amer. J. Hum. Genet.* **81** 559–575.

SUN, L., WANG, Z., LU, T., MANOLIO, T. A. and PATERSON, A. D. (2023). eXclusionarY: 10 years later, where are the sex chromosomes in GWASs? *Amer. J. Hum. Genet.* **110** 903–912.

TALIUN, D., HARRIS, D. N., KESSLER, M. D., CARLSON, J., SZPIECH, Z. A., TORRES, R., TALIUN, S. A. G., CORVELO, A., GOGARTEN, S. M. et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature* **590** 290–299.

THE 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74. https://doi.org/10.1038/nature15393

WANG, Z., PATERSON, A. D. and SUN, L. (2024). Supplement to "A population-aware retrospective regression to detect genome-wide variants with sex difference in allele frequency." https://doi.org/10.1214/23-AOAS1825SUPPA, https://doi.org/10.1214/23-AOAS1825SUPPB, https://doi.org/10.1214/23-AOAS1825SUPPC

WANG, Z., SUN, L. and PATERSON, A. D. (2022). Major sex differences in allele frequencies for X chromosomal variants in both the 1000 Genomes Project and gnomAD. *PLoS Genet.* **18** e1010231. https://doi.org/10.1371/journal.pgen.1010231

WILLER, C. J., LI, Y. and ABECASIS, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26** 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

WISE, A. L., GYI, L. and MANOLIO, T. A. (2013). eXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Amer. J. Hum. Genet.* **92** 643–647. https://doi.org/10.1016/j.ajhg.2013.03.017

YE, T., LIU, Z., SUN, B. and TCHETGEN, E. T. (2021). GENIUS-MAWII: For robust Mendelian randomization with many weak invalid instruments. arXiv preprint. Available at: arXiv:2107.06238.

ZHANG, L. and SUN, L. (2022a). A generalized robust allele-based genetic association test. *Biometrics* **78** 487–498. MR4450570 https://doi.org/10.1111/biom.13456

ZHANG, L. and SUN, L. (2022b). Unifying genetic association tests via regression: Prospective and retrospective, parametric and nonparametric, and genotype- and allele-based tests. *Canad. J. Statist.* **50** 1321–1338. MR4521919

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1137–1159
https://doi.org/10.1214/23-AOAS1826

BAYESIAN NESTED LATENT CLASS MODELS FOR CAUSE-OF-DEATH ASSIGNMENT USING VERBAL AUTOPSIES ACROSS MULTIPLE DOMAINS

BY ZEHANG RICHARD LI[1,a], ZHENKE WU[2,b], IRENA CHEN[3,c] AND SAMUEL J. CLARK[4,d]

[1]*Department of Statistics, University of California, Santa Cruz,* [a]*lizehang@ucsc.edu*
[2]*Department of Biostatistics, University of Michigan,* [b]*zhenkewu@umich.edu*
[3]*Department of Digital and Computational Demography, Max Planck Institute for Demographic Research,* [c]*chen@demogr.mpg.de*
[4]*Department of Sociology, The Ohio State University,* [d]*work@samclark.net*

Understanding cause-specific mortality rates is crucial for monitoring population health and designing public health interventions. Worldwide, two-thirds of deaths do not have a cause assigned. Verbal autopsy (VA) is a well-established tool to collect information describing deaths outside of hospitals by conducting surveys to caregivers of a deceased person. It is routinely implemented in many low- and middle-income countries. Statistical algorithms to assign cause of death using VAs are typically vulnerable to the distribution shift between the data used to train the model and the target population. This presents a major challenge for analyzing VAs, as labeled data are usually unavailable in the target population. This article proposes a latent class model framework for VA data (LCVA) that jointly models VAs collected over multiple heterogeneous domains, assigns causes of death for out-of-domain observations and estimates cause-specific mortality fractions for a new domain. We introduce a parsimonious representation of the joint distribution of the collected symptoms using nested latent class models and develop a computationally efficient algorithm for posterior inference. We demonstrate that LCVA outperforms existing methods in predictive performance and scalability. Supplementary Material and reproducible analysis codes are available online. The R package LCVA implementing the method is available on GitHub (https://github.com/richardli/LCVA).

REFERENCES

ACKERMAN, B., SIDDIQUE, J. and STUART, E. A. (2019). Transportability of outcome measurement error correction: From validation studies to intervention trials. arXiv preprint. Available at arXiv:1907.10722.

ADAMS, R. P. and GHAHRAMANI, Z. (2009). Archipelago: Nonparametric Bayesian semi-supervised learning. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 1–8. MR4172311

BLANCO, A., PEREZ, A., CASILLAS, A. and COBOS, D. (2020). Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE J. Biomed. Health Inform.*

BOLUKI, S., QIAN, X. and DOUGHERTY, E. R. (2021). Optimal Bayesian supervised domain adaptation for RNA sequencing data. *Bioinformatics* **37** 3212–3219. https://doi.org/10.1093/bioinformatics/btab228

BREIMAN, R. F., BLAU, D. M., MUTEVEDZI, P., AKELO, V., MANDOMANDO, I., OGBUANU, I. U., SOW, S. O., MADRID, L., EL ARIFEEN, S. et al. (2021). Postmortem investigations and identification of multiple causes of child deaths: An analysis of findings from the Child Health and Mortality Prevention Surveillance (CHAMPS) network. *PLoS Med.* **18** e1003814.

BRUZZONE, L. and PRIETO, D. F. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **39** 456–460.

BYASS, P., CHANDRAMOHAN, D., CLARK, S. J., D'AMBRUOSO, L., FOTTRELL, E., GRAHAM, W. J., HERBST, A. J., HODGSON, A., HOUNTON, S. et al. (2012). Strengthening standardised interpretation of verbal autopsy data: The new InterVA-4 tool. *Glob. Health Action* **5**.

Key words and phrases. Domain adaptation, data shift, mixture model, dependent binary data, quantification learning.

BYASS, P., HUONG, D. L. and VAN MINH, H. (2003). A probabilistic approach to interpreting verbal autopsies: Methodology and preliminary validation in Vietnam. *Scand. J. Soc. Health* **31** 32–37.
BYASS, P., HUSSAIN-ALKHATEEB, L., D'AMBRUOSO, L., CLARK, S., DAVIES, J., FOTTRELL, E., BIRD, J., KABUDULA, C., TOLLMAN, S. et al. (2019). An integrated approach to processing WHO-2016 verbal autopsy data: The InterVA-5 model. *BMC Med.* **17** 1–12.
CLARK, S. J., LI, Z. R. and MCCORMICK, T. H. (2018). Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. Available at arXiv:1803.07141.
CUCALA, L., MARIN, J.-M., ROBERT, C. P. and TITTERINGTON, D. M. (2009). A Bayesian reassessment of nearest-neighbor classification. *J. Amer. Statist. Assoc.* **104** 263–273. MR2663042 https://doi.org/10.1198/jasa.2009.0125
DATTA, A., FIKSEL, J., AMOUZOU, A. and ZEGER, S. L. (2021). Regularized Bayesian transfer learning for population-level etiological distributions. *Biostatistics* **22** 836–857. MR4325730 https://doi.org/10.1093/biostatistics/kxaa001
DAUMÉ, H. III and MARCU, D. (2006). Domain adaptation for statistical classifiers. *J. Artificial Intelligence Res.* **26** 101–126. MR2306416 https://doi.org/10.1613/jair.1872
DE VITO, R., BELLIO, R., TRIPPA, L. and PARMIGIANI, G. (2019). Multi-study factor analysis. *Biometrics* **75** 337–346. MR3953734 https://doi.org/10.1111/biom.12974
DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004 https://doi.org/10.1198/jasa.2009.tm08439
FIKSEL, J., DATTA, A., AMOUZOU, A. and ZEGER, S. (2021). Generalized Bayes quantification learning under dataset shift. *J. Amer. Statist. Assoc.* **117** 2163–2181. MR4528496 https://doi.org/10.1080/01621459.2021.1909599
FLAXMAN, A. D., VAHDATPOUR, A., GREEN, S., JAMES, S. L. and MURRAY, C. J. (2011). Random forests for verbal autopsy analysis: Multisite validation study using clinical diagnostic gold standards. *Popul. Health Metr.* **9** 29. https://doi.org/10.1186/1478-7954-9-29
GLOROT, X., BORDES, A. and BENGIO, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 513–520.
GONZÁLEZ, P., CASTAÑO, A., CHAWLA, N. V. and COZ, J. J. D. (2017). A review on quantification learning. *ACM Comput. Surv.* **50** 1–40.
GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936 https://doi.org/10.1093/biomet/61.2.215
HAJIRAMEZANALI, E., ZAMANI DADANEH, S., KARBALAYGHAREH, A., ZHOU, M. and QIAN, X. (2018). Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Adv. Neural Inf. Process. Syst.* **31**.
JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules. arXiv preprint. Available at arXiv:1708.08719.
KING, G. and LU, Y. (2008). Verbal autopsy methods with multiple causes of death. *Statist. Sci.* **23** 78–91. MR2523943 https://doi.org/10.1214/07-STS247
KUNIHAMA, T., LI, Z. R., CLARK, S. J. and MCCORMICK, T. H. (2020). Bayesian factor models for probabilistic cause of death assessment with verbal autopsies. *Ann. Appl. Stat.* **14** 241–256. MR4085092 https://doi.org/10.1214/19-AOAS1253
LAPARRA, E., BETHARD, S. and MILLER, T. A. (2020). Rethinking domain adaptation for machine learning over clinical language. *J. Amer. Med. Inform. Assoc.* **3** 146–150. https://doi.org/10.1093/jamiaopen/ooaa010
LI, Z. R., MCCOMICK, T. H. and CLARK, S. J. (2020). Using Bayesian latent Gaussian graphical models to infer symptom associations in verbal autopsies. *Bayesian Anal.* **15** 781–807. MR4132650 https://doi.org/10.1214/19-BA1172
LI, Z. R., THOMAS, J., CHOI, E., MCCORMICK, T. H. and CLARK, S. J. (2023). The openVA toolkit for verbal autopsies. *R J.* 316–334.
LI, Z. R., WU, Z., CHEN, I. and CLARK, S. J. (2024). Supplement to "Bayesian nested latent class models for cause-of-death assignment using verbal autopsies across multiple domains." https://doi.org/10.1214/23-AOAS1826SUPPA, https://doi.org/10.1214/23-AOAS1826SUPPB
LIU, Y. (2021). Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning* 6725–6735. PMLR.
LIU, Y. and GUO, H. (2020). Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning* 6226–6236. PMLR.
LOZANO, R., LOPEZ, A. D., ATKINSON, C., NAGHAVI, M., FLAXMAN, A. D. and MURRAY, C. J. L. (2011). Performance of physician-certified verbal autopsies: Multisite validation study using clinical diagnostic gold standards. *Popul. Health Metr.* **9** 1–13.

LUNN, D., BEST, N., SPIEGELHALTER, D., GRAHAM, G. and NEUENSCHWANDER, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *J. Pharmacokinet. Pharmacodyn.* **36** 19–38. https://doi.org/10.1007/s10928-008-9109-1

MAHER, D., BIRARO, S., HOSEGOOD, V. and ISINGO, R. (2010). Translating global health research aims into action: The example of the ALPHA network. *TM IH, Trop. Med. Int. Health* **15** 321–328.

MCCORMICK, T. H., LI, Z. R., CALVERT, C., CRAMPIN, A. C., KAHN, K. and CLARK, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *J. Amer. Statist. Assoc.* **111** 1036–1049. MR3561927 https://doi.org/10.1080/01621459.2016.1152191

MHASAWADE, V., REHMAN, N. A. and CHUNARA, R. (2020). Population-aware hierarchical Bayesian domain adaptation via multi-component invariant learning. In *Proceedings of the ACM Conference on Health, Inference, and Learning* 182–192.

MIASNIKOF, P., GIANNAKEAS, V., GOMES, M., ALEKSANDROWICZ, L., SHESTOPALOFF, A. Y., ALAM, D., TOLLMAN, S., SAMARIKHALAJ, A. and JHA, P. (2015). Naive Bayes classifiers for verbal autopsies: Comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med.* **13** 1.

MORAN, K. R., TURNER, E. L., DUNSON, D. and HERRING, A. H. (2021). Bayesian hierarchical factor regression models to infer cause of death from verbal autopsy data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 532–557. MR4275835 https://doi.org/10.1111/rssc.12468

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

MUANDET, K., BALDUZZI, D. and SCHÖLKOPF, B. (2013). Domain generalization via invariant feature representation. In *International Conference on Machine Learning* 10–18. PMLR, Atlanta, GA, USA.

MURRAY, C. J. L., LOPEZ, A. D., BLACK, R., AHUJA, R., ALI, S. M., BAQUI, A., DANDONA, L., DANTZER, E., DAS, V. et al. (2011a). Population Health Metrics Research Consortium gold standard verbal autopsy validation study: Design, implementation, and development of analysis datasets. *Popul. Health Metr.* **9** 27.

MURRAY, C. J. L., LOZANO, R., FLAXMAN, A. D., VAHDATPOUR, A. and LOPEZ, A. D. (2011b). Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popul. Health Metr.* **9** 28.

NKENGASONG, J., GUDO, E., MACICAME, I., MAUNZE, X., AMOUZOU, A., BANKE, K., DOWELL, S. and JANI, I. (2020). Improving birth and death data for African decision making. *Lancet Glob. Health* **8** e35–e36. https://doi.org/10.1016/S2214-109X(19)30397-3

OQUAB, M., BOTTOU, L., LAPTEV, I. and SIVIC, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1717–1724.

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

POMPE, E. and JACOB, P. E. (2021). Asymptotics of cut distributions and robust modular inference using posterior bootstrap. arXiv preprint. Available at arXiv:2110.11149.

RAGHURAM, J., MILLER, D. J. and KESIDIS, G. (2012). Semisupervised domain adaptation for mixture model based classifiers. In 2012 46*th Annual Conference on Information Sciences and Systems* (*CISS*) 1–6. IEEE, Princeton NJ USA.

RAMPONI, A. and PLANK, B. (2020). Neural unsupervised domain adaptation in NLP—a survey. arXiv preprint. Available at arXiv:2006.00632.

REHMAN, N. A., ALIAPOULIOS, M. M., UMARWANI, D. and CHUNARA, R. (2018). Domain adaptation for infection prediction from symptoms based on data from different study designs and contexts. arXiv preprint. Available at arXiv:1806.08835.

SANKOH, O. and BYASS, P. (2012). The INDEPTH network: Filling vital gaps in global epidemiology. *Int. J. Epidemiol.* **41** 579–588. https://doi.org/10.1093/ije/dys081

SCHÖLKOPF, B., JANZING, D., PETERS, J., SGOURITSA, E., ZHANG, K. and MOOIJ, J. (2012). On causal and anticausal learning. In *International Conference on Machine Learning* 459–466. PMLR, Edinburgh, Scotland, UK.

SERINA, P., RILEY, I., STEWART, A., JAMES, S. L., FLAXMAN, A. D., LOZANO, R., HERNANDEZ, B., MOONEY, M. D., LUNING, R. et al. (2015). Improving performance of the Tariff method for assigning causes of death to verbal autopsies. *BMC Med.* **13** 1.

SHEN, K., JONES, R., KUMAR, A., XIE, S. M., HAOCHEN, J. Z., MA, T. and LIANG, P. (2022). Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. arXiv preprint. Available at arXiv:2204.00570.

SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4

STORKEY, A. (2009). When training and test sets are different: Characterizing learning transfer. In *Dataset Shift in Machine Learning* **30** 3–28.

TZENG, E., HOFFMAN, J., ZHANG, N., SAENKO, K. and DARRELL, T. (2014). Deep domain confusion: Maximizing for domain invariance. arXiv preprint. Available at arXiv:1412.3474.

WANG, M. and DENG, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing* **312** 135–153.

WILSON, G. and COOK, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* **11** 1–46.

WOOD, F. and TEH, Y. W. (2009). A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Artificial Intelligence and Statistics* 607–614. PMLR, Clearwater Beach, FL, USA.

WORLD HEALTH ORGANIZATION (2021). WHO civil registration and vital statistics strategic implementation plan 2021–2025.

WU, X., BRAUN, D., KIOUMOURTZOGLOU, M.-A., CHOIRAT, C., DI, Q. and DOMINICI, F. (2019). Causal inference in the context of an error prone exposure: Air pollution and mortality. *Ann. Appl. Stat.* **13** 520–547. MR3937439 https://doi.org/10.1214/18-AOAS1206

WU, Z., LI, Z. R., CHEN, I. and LI, M. (2021). Tree-informed Bayesian multi-source domain adaptation: Cross-population probabilistic cause-of-death assignment using verbal autopsy. arXiv preprint. Available at arXiv:2112.10978.

YAO, Y., VEHTARI, A. and GELMAN, A. (2022). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *J. Mach. Learn. Res.* **23** Paper No. 79, 45. MR4576664

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

ZIGLER, C. M., WATTS, K., YEH, R. W., WANG, Y., COULL, B. A. and DOMINICI, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics* **69** 263–273. MR3058073 https://doi.org/10.1111/j.1541-0420.2012.01830.x

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1160–1177
https://doi.org/10.1214/23-AOAS1827

FILTRATED COMMON FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS OF MULTIGROUP FUNCTIONAL DATA

BY SHUHAO JIAO[1,a], RON FROSTIG[2,b] AND HERNANDO OMBAO[3,c]

[1]*Department of Biostatistics, City University of Hong Kong,* [a]*shuhao.jiao@cityu.edu.hk*
[2]*Department of Neurobiology and Behavior, University or California, Irvine,* [b]*rfrostig@uci.edu*
[3]*Statistics Program, King Abdullah University of Science and Technology,* [c]*hernando.ombao@kaust.edu.sa*

Local field potentials (LFPs) are signals that measure electrical activities in localized cortical regions and are collected from multiple tetrodes implanted across a patch on the surface of cortex. Hence, they can be treated as multigroup functional data, where the trajectories collected across temporal epochs from one tetrode are viewed as a group of functions. In many cases multitetrode LFP trajectories contain both global variation patterns (which are shared by most groups, due to signal synchrony) and idiosyncratic variation patterns (common only to a small subset of groups), and such structure is very informative to the data mechanism. Therefore, one goal in this paper is to develop an efficient algorithm that is able to capture and quantify both global and idiosyncratic features. We develop the novel filtrated common functional principal components (filt-fPCA) method, which is a novel forest-structured fPCA for multigroup functional data. A major advantage of the proposed filt-fPCA method is its ability to extract the common components in a flexible "multiresolution" manner. The proposed approach is highly data-driven, and no prior knowledge of "ground-truth" data structure is needed, making it suitable for analyzing complex multigroup functional data. In addition, the filt-fPCA method is able to produce parsimonious, interpretable, and efficient functional reconstruction (low reconstruction error) for multigroup functional data with orthonormal basis functions. Here the proposed filt-fPCA method is employed to study the impact of a shock (induced stroke) on the synchrony structure of rat brain. The proposed filt-fPCA is general and inclusive that can be readily applied to analyze any multigroup functional data, such as multivariate functional data, spatial-temporal data, and longitudinal functional data.

REFERENCES

BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J.-L. (2011). Robust functional principal components: A projection-pursuit approach. *Ann. Statist.* **39** 2852–2882. MR3012394 https://doi.org/10.1214/11-AOS923

BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343 https://doi.org/10.1214/07-AOS516

BERRENDERO, J. R., JUSTEL, A. and SVARC, M. (2011). Principal components for multivariate functional data. *Comput. Statist. Data Anal.* **55** 2619–2634. MR2802340 https://doi.org/10.1016/j.csda.2011.03.011

CHEN, K., DELICADO, P. and MÜLLER, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 177–196. MR3597969 https://doi.org/10.1111/rssb.12160

CHEN, K. and MÜLLER, H.-G. (2012). Modeling repeated functional observations. *J. Amer. Statist. Assoc.* **107** 1599–1609. MR3036419 https://doi.org/10.1080/01621459.2012.734196

CHIOU, J.-M., CHEN, Y.-T. and YANG, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statist. Sinica* **24** 1571–1596. MR3308652

COFFEY, N., HARRISON, A. J., DONOGHUE, O. A. and HAYES, K. (2011). Common functional principal components analysis: A new approach to analyzing human movement data. *Hum. Mov. Sci.* **30** 1144–1166. https://doi.org/10.1016/j.humov.2010.11.005

Key words and phrases. Functional principal components, community detection, dimension reduction, multigroup functional data, network filtration, weighted network.

CRAINICEANU, C. M., CAFFO, B. S., LUO, S., ZIPUNNIKOV, V. M. and PUNJABI, N. M. (2011). Population value decomposition, a framework for the analysis of image populations. *J. Amer. Statist. Assoc.* **106** 775–790. MR2894733 https://doi.org/10.1198/jasa.2011.ap10089

DI, C., CRAINICEANU, C. M. and JANK, W. S. (2014). Multilevel sparse functional principal component analysis. *Stat* **3** 126–143. MR4027332 https://doi.org/10.1002/sta4.50

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

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

FLURY, B. K. (1987). Two generalizations of the common principal component model. *Biometrika* **74** 59–69. MR0885919 https://doi.org/10.1093/biomet/74.1.59

FLURY, B. N. (1984). Common principal components in k groups. *J. Amer. Statist. Assoc.* **79** 892–898. MR0770284

GREVEN, S., CRAINICEANU, C., CAFFO, B. and REICH, D. (2010). Longitudinal functional principal component analysis. *Electron. J. Stat.* **4** 1022–1054. MR2727452 https://doi.org/10.1214/10-EJS575

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 109–126. MR2212577 https://doi.org/10.1111/j.1467-9868.2005.00535.x

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365 https://doi.org/10.1214/009053606000000272

HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc.* **113** 649–659. MR3832216 https://doi.org/10.1080/01621459.2016.1273115

JACQUES, J. and PREDA, C. (2014). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

JIANG, C.-R. and WANG, J.-L. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *Ann. Statist.* **38** 1194–1226. MR2604710 https://doi.org/10.1214/09-AOS742

JIAO, S., FROSTIG, R. D. and OMBAO, H. (2023). Break point detection for functional covariance. *Scand. J. Stat.* **50** 477–512. MR4599922 https://doi.org/10.1111/sjos.12589

JIAO, S., FROSTIG, R. and OMBAO, H. (2024). Supplement to "Filtrated common functional principal component analysis of multigroup functional data." https://doi.org/10.1214/23-AOAS1827SUPP

KAYANO, M. and KONISHI, S. (2009). Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data. *J. Statist. Plann. Inference* **139** 2388–2398. MR2508000 https://doi.org/10.1016/j.jspi.2008.11.002

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542. MR3086429 https://doi.org/10.1214/12-AOAS597

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765. MR0740928 https://doi.org/10.1214/aos/1176346522

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

SCHOTT, J. R. (1999). Partial common principal component subspaces. *Biometrika* **86** 899–908. MR1741985 https://doi.org/10.1093/biomet/86.4.899

WANG, B., LUO, X., ZHAO, Y. and CAFFO, B. (2021). Semiparametric partial common principal component analysis for covariance matrices. *Biometrics* **77** 1175–1186. MR4357829 https://doi.org/10.1111/biom.13369

WANN, E. G. (2017). Large-scale spatiotemporal neuronal activity dynamics predict cortical viability in a rodent model of ischemic stroke. Ph.D. thesis, UC Irvine.

YAO, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statist. Sinica* **17** 965–983. MR2408647

YAO, F. and LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 3–25. MR2212572 https://doi.org/10.1111/j.1467-9868.2005.00530.x

ZHANG, Y., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. With supplementary material available online. MR2656055 https://doi.org/10.1198/jasa.2009.tm08013

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1178–1194
https://doi.org/10.1214/23-AOAS1829

ACCURATE ESTIMATION OF RARE CELL-TYPE FRACTIONS FROM TISSUE OMICS DATA VIA HIERARCHICAL DECONVOLUTION

BY PENGHUI HUANG[1,a], MANQI CAI[1,b], XINGHUA LU[2,d], CHRIS MCKENNAN[3,e] AND JIEBIAO WANG[1,c]

[1]*Department of Biostatistics, University of Pittsburgh,* [a]*huangpenghui@pitt.edu,* [b]*mac538@pitt.edu,* [c]*jbwang@pitt.edu*
[2]*Department of Biomedical Informatics, University of Pittsburgh,* [d]*xinghua@pitt.edu*
[3]*Department of Statistics, University of Pittsburgh,* [e]*chm195@pitt.edu*

Bulk transcriptomics in tissue samples reflects the average expression levels across different cell types and is highly influenced by cellular fractions. As such, it is critical to estimate cellular fractions to both deconfound differential expression analyses and infer cell type-specific differential expression. Since experimentally counting cells is infeasible in most tissues and studies, *in silico* cellular deconvolution methods have been developed as an alternative. However, existing methods are designed for tissues consisting of clearly distinguishable cell types and have difficulties estimating highly correlated or rare cell types. To address this challenge, we propose hierarchical deconvolution (HiDecon) that uses single-cell RNA sequencing references and a hierarchical cell-type tree, which models the similarities among cell types and cell differentiation relationships, to estimate cellular fractions in bulk data. By coordinating cell fractions across layers of the hierarchical tree, cellular fraction information is passed up and down the tree, which helps correct estimation biases by pooling information across related cell types. The flexible hierarchical tree structure also enables estimating rare cell fractions by splitting the tree to higher resolutions. Through simulations and real data applications with the ground truth of measured cellular fractions, we demonstrate that HiDecon outperforms existing methods and accurately estimates cellular fractions. Finally, we show the utility of HiDecon estimates in identifying the associations between cellular fractions and Alzheimer's disease.

REFERENCES

AVILA COBOS, F., ALQUICIRA-HERNANDEZ, J., POWELL, J. E., MESTDAGH, P. and DE PRETER, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11** 1–14.

BERGER, R. L. (1997). Likelihood ratio tests and intersection-union tests. In *Advances in Statistical Decision Theory and Applications. Stat. Ind. Technol.* 225–237. Birkhäuser, Boston, MA. MR1479187

CAI, M., YUE, M., CHEN, T., LIU, J., FORNO, E., LU, X., BILLIAR, T. and CELEDÓN, J. (2022). Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution. *Bioinformatics* **38** 3004–3010.

CHEN, L., LI, Z. and WU, H. (2023). CeDAR: Incorporating cell type hierarchy improves cell type-specific differential analyses in bulk omics data. *Genome Biol.* **24** 37.

CHEN, S., WANG, J., CICEK, E. and ROEDER, K. (2020). De novo missense variants disrupting protein–protein interactions affect risk for autism through gene co-expression and protein networks in neuronal cell types. *Mol. Autism* **11** 1–16.

DAWBER, T. R., MEADORS, G. F. and MOORE JR, F. E. (1951). Epidemiological approaches to heart disease: The Framingham study. *Amer. J. Public Health Nation's Health* **41** 279–286.

FEINLEIB, M., KANNEL, W. B., GARRISON, R. J., MCNAMARA, P. M. and CASTELLI, W. P. (1975). The Framingham offspring study. Design and preliminary data. *Prev. Med.* **4** 518–525.

FISCHER, S. and GILLIS, J. (2021). How many markers are needed to robustly determine a cell's type? *iScience* **24** 103292. https://doi.org/10.1016/j.isci.2021.103292

Key words and phrases. Cellular deconvolution, single-cell data, RNA sequencing, hierarchical tree, penalized regression.

HANSEN, D. V., HANSON, J. E. and SHENG, M. (2018). Microglia in Alzheimer's disease. *J. Cell Biol.* **217** 459–472. https://doi.org/10.1083/jcb.201709069

HODGE, R. D., BAKKEN, T. E., MILLER, J. A., SMITH, K. A., BARKAN, E. R., GRAYBUCK, L. T., CLOSE, J. L., LONG, B., JOHANSEN, J. N. et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573** 61–68.

HUANG, P., CAI, M., LU, X., MCKENNAN, C. and WANG, J. (2024). Supplement to "Accurate estimation of rare cell-type fractions from tissue omics data via hierarchical deconvolution." https://doi.org/10.1214/23-AOAS1829SUPPA, https://doi.org/10.1214/23-AOAS1829SUPPB

HUNT, G. J., FREYTAG, S., BAHLO, M. and GAGNON-BARTSCH, J. A. (2019). dtangle: Accurate and robust cell type deconvolution. *Bioinformatics* **35** 2093–2099. https://doi.org/10.1093/bioinformatics/bty926

JAFFE, A. E. and IRIZARRY, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15** R31. https://doi.org/10.1186/gb-2014-15-2-r31

JIA, C., HU, Y., KELLY, D., KIM, J., LI, M. and ZHANG, N. R. (2017). Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.* **45** 10978–10988.

JIN, C., CHEN, M., LIN, D.-Y. and SUN, W. (2021). Cell-type-aware analysis of RNA-seq data. *Nat. Comput. Sci.* **1** 253–261.

LI, Z., WU, Z., JIN, P. and WU, H. (2019). Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics* **35** 3898–3905.

LIN, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 255–268.

MATHYS, H., DAVILA-VELDERRAIN, J., PENG, Z., GAO, F., MOHAMMADI, S., YOUNG, J. Z., MENON, M., HE, L., ABDURROB, F. et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570** 332–337.

MILLER, J. A., GOUWENS, N. W., TASIC, B., COLLMAN, F., VAN VELTHOVEN, C. T., BAKKEN, T. E., HAWRYLYCZ, M. J., ZENG, H., LEIN, E. S. et al. (2020). Common cell type nomenclature for the mammalian brain. *eLife* **9**. https://doi.org/10.7554/eLife.59928

MOHAMMADI, S., ZUCKERMAN, N., GOLDSMITH, A. and GRAMA, A. (2016). A critical survey of deconvolution methods for separating cell types in complex tissues. *Proc. IEEE* **105** 340–366.

MOSTAFAVI, S., GAITERI, C., SULLIVAN, S. E., WHITE, C. C., TASAKI, S., XU, J., TAGA, M., KLEIN, H.-U., PATRICK, E. et al. (2018). A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21** 811–819.

NEWMAN, A. M., LIU, C. L., GREEN, M. R., GENTLES, A. J., FENG, W., XU, Y., HOANG, C. D., DIEHN, M. and ALIZADEH, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12** 453–457. https://doi.org/10.1038/nmeth.3337

PENG, M., WAMSLEY, B., ELKINS, A. G., GESCHWIND, D. H., WEI, Y. and ROEDER, K. (2021). Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree. *Nucleic Acids Res.* **49** e91–e91.

REN, X., WEN, W., FAN, X., HOU, W., SU, B., CAI, P., LI, J., LIU, Y., TANG, F. et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184** 1895–1913.

SPLANSKY, G. L., COREY, D., YANG, Q., ATWOOD, L. D., CUPPLES, L. A., BENJAMIN, E. J., D'AGOSTINO SR, R. B., FOX, C. S., LARSON, M. G. et al. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham heart study: Design, recruitment, and initial examination. *Amer. J. Epidemiol.* **165** 1328–1335.

WANG, J., DEVLIN, B. and ROEDER, K. (2020). Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics* **36** 782–788. https://doi.org/10.1093/bioinformatics/btz619

WANG, J., ROEDER, K. and DEVLIN, B. (2021). Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* **31** 1807–1818. https://doi.org/10.1101/gr.268722.120

WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R. and LI, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10** 1–9.

WESTRA, H.-J., ARENDS, D., ESKO, T., PETERS, M. J., SCHURMANN, C., SCHRAMM, K., KETTUNEN, J., YAGHOOTKAR, H., FAIRFAX, B. P. et al. (2015). Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11** e1005223.

WILSON, D. R., JIN, C., IBRAHIM, J. G. and SUN, W. (2020). ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *J. Amer. Statist. Assoc.* **115** 1055–1065. MR4143449 https://doi.org/10.1080/01621459.2019.1654874

WU, Z. and WU, H. (2020). Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. *Genome Biol.* **21**.

ZHENG, S. C., WEBSTER, A. P., DONG, D., FEBER, A., GRAHAM, D. G., SULLIVAN, R., JEVONS, S., LOVAT, L. B., BECK, S. et al. (2018). A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics* **10** 925–940.

ZHONG, Y., WAN, Y., PANG, K. CHOW, L. M. L. and LIU, Z. (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinform.* **14** 1–10.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1195–1212
https://doi.org/10.1214/23-AOAS1830

TENSOR REGRESSION FOR INCOMPLETE OBSERVATIONS WITH APPLICATION TO LONGITUDINAL STUDIES

BY TIANCHEN XU[1,a], KUN CHEN[2,b] AND GEN LI[3,c]

[1]*Bristol Myers Squibb,* [a]*tx2155@columbia.edu*
[2]*Department of Statistics, University of Connecticut,* [b]*kun.chen@uconn.edu*
[3]*Department of Biostatistics, University of Michigan, Ann Arbor,* [c]*ligen@umich.edu*

Multivariate longitudinal data are frequently encountered in practice such as in our motivating longitudinal microbiome study. It is of general interest to associate such high-dimensional, longitudinal measures with some univariate continuous outcome. However, incomplete observations are common in a regular study design, as not all samples are measured at every time point, giving rise to the so-called blockwise missing values. Such missing structure imposes significant challenges for association analysis and defies many existing methods that require complete samples. In this paper we propose to represent multivariate longitudinal data as a three-way tensor array (i.e., sample-by-feature-by-time) and exploit a parsimonious scalar-on-tensor regression model for association analysis. We develop a regularized covariance-based estimation procedure that effectively leverages all available observations without imputation. The method achieves variable selection and smooth estimation of time-varying effects. The application to the motivating microbiome study reveals interesting links between the preterm infant's gut microbiome dynamics and their neurodevelopment. Additional numerical studies on synthetic data and a longitudinal aging study further demonstrate the efficacy of the proposed method.

REFERENCES

AATSINKI, A.-K., LAHTI, L., UUSITUPA, H.-M., MUNUKKA, E., KESKITALO, A., NOLVI, S., O'MAHONY, S., PIETILÄ, S., ELO, L. L. et al. (2019). Gut microbiota composition is associated with temperament traits in infants. *Brain Behav. Immun.* **80** 849–858.

BIJMA, F., DE MUNCK, J. C. and HEETHAAR, R. M. (2005). The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. *NeuroImage* **27** 402–415. https://doi.org/10.1016/j.neuroimage.2005.04.015

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876

CHERNOZHUKOV, V., HANSEN, C., LIAO, Y. and ZHU, Y. (2023). Inference for low-rank models. *Ann. Statist.* **51** 1309–1330. MR4630950 https://doi.org/10.1214/23-aos2293

CONG, X., JUDGE, M., XU, W., DIALLO, A., JANTON, S., BROWNELL, E. A., MAAS, K. and GRAF, J. (2017). Influence of feeding type on gut microbiome development in hospitalized preterm infants. *Nursing Res.* **66** 123–133. https://doi.org/10.1097/NNR.0000000000000208

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

DINAN, T. G. and CRYAN, J. F. (2012). Regulation of the stress response by the gut microbiota: Implications for psychoneuroendocrinology. *Psychoneuroendocrinology* **37** 1369–1378. https://doi.org/10.1016/j.psyneuen.2012.03.007

DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.* **64** 105–123.

ENDERS, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabil. Psychol.* **56** 267–288. https://doi.org/10.1037/a0025579

Key words and phrases. Scalar-on-tensor regression, blockwise missing data, longitudinal measurements, microbiome data.

GERBER, G. K. (2015). Longitudinal microbiome data analysis. In *Metagenomics for Microbiology* 97–111. Elsevier, New York.

GLANZ, H. and CARVALHO, L. (2018). An expectation-maximization algorithm for the matrix normal distribution with an application in remote sensing. *J. Multivariate Anal.* **167** 31–48. MR3830632 https://doi.org/10.1016/j.jmva.2018.03.010

GLOOR, G. B., WU, J. R., PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2016). It's all relative: Analyzing microbiome data as compositions. *Ann. Epidemiol.* **26** 322–329. https://doi.org/10.1016/j.annepidem.2016.03.003

GOH, G., DEY, D. K. and CHEN, K. (2017). Bayesian sparse reduced rank multivariate regression. *J. Multivariate Anal.* **157** 14–28. MR3641733 https://doi.org/10.1016/j.jmva.2017.02.007

GOTTFRIDSSON, A. (2011). Likelihood ratio tests of separable or double separable covariance structure, and the empirical null distribution.

GU, X. and MATLOFF, N. (2015). A different approach to the problem of missing data. arXiv preprint. Available at arXiv:1509.04992.

HAN, R., SHI, P. and ZHANG, A. R. (2021). Guaranteed functional tensor singular value decomposition. arXiv preprint. Available at arXiv:2108.04201.

HARSHMAN, R. A. et al. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis.

HE, L., CHEN, K., XU, W., ZHOU, J. and WANG, F. (2018). Boosted sparse and low-rank tensor regression. *Adv. Neural Inf. Process. Syst.* **31**.

HIGHAM, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA J. Numer. Anal.* **22** 329–343. MR1918653 https://doi.org/10.1093/imanum/22.3.329

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238 https://doi.org/10.1214/11-BA606

HUANG, J. Z., SHEN, H., BUJA, A. et al. (2008). Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2** 678–695. MR2426107 https://doi.org/10.1214/08-EJS218

HUSSAIN, S., HUSSAIN, S. and ASHRAF, M. (2018). Pneumonia and bacteraemia caused by Gemella morbillorum in a previously healthy infant: First reported case in literature. *Case Rep.* **2018** bcr–2018.

JIANG, L. (2020). *Statistical Methods for Longitudinal Data Analysis and Reproducible Feature Selection in Human Microbiome Studies*. Univ. California, San Diego, CA.

JIANG, L., ELROD, C., KIM, J. J., SWAFFORD, A. D., KNIGHT, R. and THOMPSON, W. K. (2022). Bayesian multivariate sparse functional principal components analysis with application to longitudinal microbiome multiomics data. *Ann. Appl. Stat.* **16** 2231–2249. MR4489207 https://doi.org/10.1214/21-aoas1587

JIANG, L., ZHONG, Y., ELROD, C., NATARAJAN, L. and KNIGHT, R. (2020). BayesTime: Bayesian functional principal components for sparse longitudinal data. arXiv preprint. Available at arXiv:2012.00579.

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

KOENIG, J. E., SPOR, A., SCALFONE, N., FRICKER, A. D., STOMBAUGH, J., KNIGHT, R., ANGENENT, L. T. and LEY, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. USA* **108** 4578–4585.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

LI, C., XIAO, L. and LUO, S. (2020). Fast covariance estimation for multivariate sparse functional data. *Stat* **9** e245. MR4116315

LI, X., XU, D., ZHOU, H. and LI, L. (2018). Tucker tensor regression and neuroimaging analysis. *Stat. Biosci.* **10** 520–545.

LIU, J., JI, S., YE, J. et al. (2009). SLEP: Sparse learning with efficient projections. Arizona State University 6 7.

LOCK, E. F. (2018). Tensor-on-tensor regression. *J. Comput. Graph. Statist.* **27** 638–647. MR3863764 https://doi.org/10.1080/10618600.2017.1401544

LU, N. and ZIMMERMAN, D. L. (2004). On likelihood-based inference for a separable covariance matrix. Tech. Rep. 337, Statistics and Actuarial Science Dept., Univ. Iowa, Iowa City, IA..

LU, N. and ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457. MR2187860 https://doi.org/10.1016/j.spl.2005.04.020

MAKALIC, E. and SCHMIDT, D. F. (2022). An efficient algorithm for sampling from $\sin^k(x)$ for generating random correlation matrices. *Comm. Statist. Simulation Comput.* **51** 2731–2735. MR4422888 https://doi.org/10.1080/03610918.2019.1700277

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957 https://doi.org/10.1016/j.cam.2012.09.017

MANI, S. and NAIR, J. (2021). Pantoea infections in the neonatal intensive care unit. *Cureus* **13** e13103. https://doi.org/10.7759/cureus.13103

MARTINO, C., SHENHAV, L. and MAROTZ, C. A. ARMSTRONG, G., MCDONALD, D., VÁZQUEZ-BAEZA, Y., MORTON, J. T., JIANG, L., DOMINGUEZ-BELLO, M. G. et al. (2021). Deconvolutes gut microbial community dynamics. *Nat. Biotechnol.* **39** 165–168.

MENG, X.-L. and RUBIN, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86** 899–909.

MWANIKI, M. K., ATIENO, M., LAWN, J. E. and NEWTON, C. R. J. C. (2012). Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: A systematic review. *Lancet* **379** 445–452.

NATIONAL CENTER FOR HEALTH STATISTICS et al. (2003). National health interview survey 1994: Second longitudinal study on aging, wave 3, 2000. US Dept. Health and Human Services, Hyattsville, MD.

PURCELL, L. K., FINLEY, J. P., CHEN, R., LOVGREN, M. and HALPERIN, S. A. (2001). Gemella species endocarditis in a child. *Can. J. Infect. Dis.* **12** 317–320.

SENGUPTA, M., BANERJEE, S., DAS, N. K., GUCHHAIT, P. and MISRA, S. (2016). Early onset neonatal septicaemia caused by Pantoea agglomerans. *J. Clin. Diagn. Res.* **10** DD01–DD02. https://doi.org/10.7860/JCDR/2016/19613.7807

SHAHIN, M., JI, B. and DIXIT, P. (2022). EMBED: Essential Microbiome Dynamics, a dimensionality reduction approach for longitudinal microbiome studies.

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **10** 1019–1040. MR3528370 https://doi.org/10.1214/16-AOAS928

SIGNORETTO, M., TRAN DINH, Q., DE LATHAUWER, L. and SUYKENS, J. A. K. (2014). Learning with tensors: A framework based on convex optimization and spectral regularization. *Mach. Learn.* **94** 303–351. MR3166226 https://doi.org/10.1007/s10994-013-5366-3

SIWAKOTI, S., SAH, R., RAJBHANDARI, R. S. and KHANAL, B. (2018). *Case Rep. Pediatr.* **2018** 4158734. https://doi.org/10.1155/2018/4158734

SORDILLO, J. E., KORRICK, S., LARANJO, N., CAREY, V., WEINSTOCK, G. M., GOLD, D. R., O'CONNOR, G., SANDEL, M., BACHARIER, L. B. et al. (2019). Association of the infant gut microbiome with early childhood neurodevelopmental outcomes: An ancillary study to the VDAART randomized clinical trial. *JAMA Netw. Open* **2** e190905–e190905.

SUN, C. and DAI, R. (2017). Rank-constrained optimization and its applications. *Automatica J. IFAC* **82** 128–136. MR3658748 https://doi.org/10.1016/j.automatica.2017.04.039

SUN, Z., XU, W., CONG, X., LI, G. and CHEN, K. (2020). Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *Ann. Appl. Stat.* **14** 1535–1556. MR4152145 https://doi.org/10.1214/20-AOAS1357

TAMANA, S. K., TUN, H. M., KONYA, T., CHARI, R. S., FIELD, C. J., GUTTMAN, D. S., BECKER, A. B., MORAES, T. J., TURVEY, S. E. et al. (2021). Bacteroides-dominant gut microbiome of late infancy is associated with enhanced neurodevelopment. *Gut Microbes* **13** 1930875.

THEOBALD, D. L. and WUTTKE, D. S. (2008). Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.* **4** e43. https://doi.org/10.1371/journal.pcbi.0040043

TIAN, W. and YUAN, X. (2016). Faster alternating direction method of multipliers with a worst-case $O(1/n^2)$ convergence rate.

TROSVIK, P., STENSETH, N. C. and RUDI, K. (2010). Convergent temporal dynamics of the human infant gut microbiota. *ISME J.* **4** 151–158.

VAN LOAN, C. F. and PITSIANIS, N. (1993). Approximation with Kronecker products. In *Linear Algebra for Large Scale and Real-Time Applications* (*Leuven*, 1992). *NATO Adv. Sci. Inst. Ser. E: Appl. Sci.* **232** 293–314. Kluwer Academic, Dordrecht. MR1250183

WERNER, K., JANSSON, M. and STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans. Signal Process.* **56** 478–491. MR2445531 https://doi.org/10.1109/TSP.2007.907834

WITTEN, D. M. and TIBSHIRANI, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 615–636. MR2749910 https://doi.org/10.1111/j.1467-9868.2009.00699.x

XIA, D., ZHANG, A. R. and ZHOU, Y. (2022). Inference for low-rank tensors—no need to debias. *Ann. Statist.* **50** 1220–1245. MR4404934 https://doi.org/10.1214/21-aos2146

XU, T., CHEN, K. and LI, G. (2022). The more data, the better? Demystifying deletion-based methods in linear regression with missing data. *Stat. Interface* **15** 515–526. MR4391876

XU, T., CHEN, K. and LI, G. (2024). Supplement to "Tensor regression for incomplete observations with application to longitudinal studies." https://doi.org/10.1214/23-AOAS1830SUPPA, https://doi.org/10.1214/23-AOAS1830SUPPB

YU, G., LI, Q., SHEN, D. and LIU, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *J. Amer. Statist. Assoc.* **115** 1406–1419. MR4143474 https://doi.org/10.1080/01621459.2019.1632079

ZHANG, M., MA, W., ZHANG, J., HE, Y. and WANG, J. (2018). Analysis of gut microbiota profiles and microbe-disease associations in children with autism spectrum disorders in China. *Sci. Rep.* **8** 1–9.

ZHOU, H. and LI, L. (2014). Regularized matrix regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 463–483. MR3164874 https://doi.org/10.1111/rssb.12031

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1213–1235
https://doi.org/10.1214/23-AOAS1831

LEARNING COMMON STRUCTURES IN A COLLECTION OF NETWORKS. AN APPLICATION TO FOOD WEBS

BY SAINT-CLAIR CHABERT-LIDDELL[a], PIERRE BARBILLON[b] AND SOPHIE DONNET[c]

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, [a]*academic@chabert-liddell.com,* [b]*pierre.barbillon@agroparistech.fr,* [c]*sophie.donnet@inrae.fr*

Let a collection of networks represent interactions within several (social or ecological) systems. We pursue two objectives: identifying similarities in the topological structures that are held in common between the networks and clustering the collection into subcollections of structurally homogeneous networks. We tackle these two questions with a probabilistic model-based approach. We propose an extension of the stochastic block model (SBM) adapted to the joint modeling of a collection of networks. The networks in the collection are assumed to be independent realizations of SBMs. The common connectivity structure is imposed through the equality of some parameters.

The model parameters are estimated with a variational expectation-maximization (EM) algorithm. We derive an ad hoc penalized likelihood criterion to select the number of blocks and to assess the adequacy of the consensus found between the structures of the different networks. This same criterion can also be used to cluster networks on the basis of their connectivity structure. It thus provides a partition of the collection into subsets of structurally homogeneous networks.

The relevance of our proposition is assessed on two collections of ecological networks. First, an application to three stream food webs reveals the homogeneity of their structures and the correspondence between groups of species in different ecosystems playing equivalent ecological roles. Moreover, the joint analysis allows a finer analysis of the structure of smaller networks. Second, we cluster 67 food webs according to their connectivity structures and demonstrate that five mesoscale structures are sufficient to describe this collection.

REFERENCES

ALLESINA, S. and PASCUAL, M. (2009). Food web models: A plea for groups. *Ecol. Lett.* **12** 652–662.

BAR-HEN, A., BARBILLON, P. and DONNET, S. (2020). Block models for generalized multipartite networks: Applications in ecology and ethnobiology. *Stat. Model.* **22** 273–296. MR4458102 https://doi.org/10.1177/1471082X20963254

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BLÜTHGEN, N., MENZEL, F. and BLÜTHGEN, N. (2006). Measuring specialization in species interaction networks. *BMC Ecol.* **6** 1–12.

BOORMAN, S. A. and WHITE, H. C. (1976). Social structure from multiple networks. II. Role structures. *Amer. J. Sociol.* **81** 1384–1446. https://doi.org/10.1086/226228

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467 https://doi.org/10.1214/12-EJS729

CHABERT-LIDDELL, S.-C., BARBILLON, P. and DONNET, S. (2022). Impact of the mesoscale structure of a bipartite ecological interaction network on its robustness through a probabilistic modeling. *Environmetrics* **33** Paper No. e2709, 20. MR4393412 https://doi.org/10.1002/env.2709

CHABERT-LIDDELL, S.-C., BARBILLON, P. and DONNET, S. (2024). Supplement to "Learning common structures in a collection of networks. An application to food webs." https://doi.org/10.1214/23-AOAS1831SUPPA, https://doi.org/10.1214/23-AOAS1831SUPPB

Key words and phrases. Stochastic block model, clustering, networks, latent variable models, ecology.

CHABERT-LIDDELL, S.-C., BARBILLON, P., DONNET, S. and LAZEGA, E. (2021). A stochastic block model approach for the analysis of multilevel networks: An application to the sociology of organizations. *Comput. Statist. Data Anal.* **158** Paper No. 107179, 25. MR4210955 https://doi.org/10.1016/j.csda.2021.107179

CHIQUET, J., DONNET, S. and BARBILLON, P. (2021). sbm: Stochastic blockmodels. R package version 0.4.3.

CIRTWILL, A. R., DALLA RIVA, G. V., GAIARSA, M. P., BIMLER, M. D., CAGUA, E. F., COUX, C. and DEHLING, D. M. (2018). A review of species role concepts in food webs. *Food Webs* **16** e00093.

CLAUSET, A., MOORE, C. and NEWMAN, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* **453** 98.

CÔME, E., JOUVIN, N., LATOUCHE, P. and BOUVEYRON, C. (2021). Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. *Adv. Data Anal. Classif.* **15** 957–986. MR4333226 https://doi.org/10.1007/s11634-021-00440-z

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817 https://doi.org/10.1007/s11222-007-9046-7

DONNAT, C. and HOLMES, S. (2018). Tracking network dynamics: A survey using graph distances. *Ann. Appl. Stat.* **12** 971–1012. MR3834292 https://doi.org/10.1214/18-AOAS1176

DURANTE, D., DUNSON, D. B. and VOGELSTEIN, J. T. (2017). Nonparametric Bayes modeling of populations of networks. *J. Amer. Statist. Assoc.* **112** 1516–1530. MR3750873 https://doi.org/10.1080/01621459.2016.1219260

FAUST, K. and SKVORETZ, J. (2002). Comparing networks across space and time, size and species. *Sociol. Method.* **32** 8, 267–299. https://doi.org/10.1111/1467-9531.00118

GOVAERT, G. and NADIF, M. (2003). Clustering with block mixture models. *Pattern Recognit.* **36** 463–473.

GUIMERÀ, R. and SALES-PARDO, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* **106** 22073–22078.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

KIVELÄ, M., ARENAS, A., BARTHELEMY, M. and GLEESON, J. P. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.

KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. *Springer Series in Statistics*. Springer, New York. MR2724362 https://doi.org/10.1007/978-0-387-88146-1

LAFFERTY, K. D., ALLESINA, S., ARIM, M. and BRIGGS, C. J. (2008). Parasites in food webs: The ultimate missing links. *Ecol. Lett.* **11** 533–546.

LE, C. M., LEVIN, K. and LEVINA, E. (2018). Estimating a network from multiple noisy realizations. *Electron. J. Stat.* **12** 4697–4740. MR3894068 https://doi.org/10.1214/18-ejs1521

LEGER, J.-B., BARBILLON, P. and CHIQUET, J. (2020). blockmodels: Latent and stochastic block model estimation by a 'V-EM' algorithm. R package version 1.1.4.

LUCZKOVICH, J. J., BORGATTI, S. P., JOHNSON, J. C. and EVERETT, M. G. (2003). Defining and measuring trophic role similarity in food webs using regular equivalence. *J. Theoret. Biol.* **220** 303–321. MR2042871 https://doi.org/10.1006/jtbi.2003.3147

MARIADASSOU, M., ROBIN, S. and VACHER, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Ann. Appl. Stat.* **4** 715–742. MR2758646 https://doi.org/10.1214/10-AOAS361

MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

MICHALSKA-SMITH, M. J. and ALLESINA, S. (2019). Telling ecological networks apart by their structure: A computational challenge. *PLoS Comput. Biol.* **15** e1007076. https://doi.org/10.1371/journal.pcbi.1007076

MUKHERJEE, S. S., SARKAR, P. and LIN, L. (2017). On clustering network-valued data. *Adv. Neural Inf. Process. Syst.* **30**.

OHLSSON, M. and EKLÖF, A. (2020). Spatial resolution and location impact group structure in a marine food web. *Ecol. Lett.* **23** 1451–1459. https://doi.org/10.1111/ele.13567

PAUL, S. and CHEN, Y. (2020). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *Ann. Appl. Stat.* **14** 993–1029. MR4117838 https://doi.org/10.1214/20-AOAS1339

PAVLOVIĆ, D. M., GUILLAUME, B. R. L., TOWLSON, E. K., KUEK, N. M. Y., AFYOUNI, S., VÉRTES, P. E., YEO, B. T. T., BULLMORE, E. T. and NICHOLS, T. E. (2020). Multi-subject stochastic blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage* **220** 116611. MR4061107 https://doi.org/10.1016/j.neuroimage.2020.116611

PEEL, L., LARREMORE, D. B. and CLAUSET, A. (2017). The ground truth about metadata and community detection in networks. *Sci. Adv.* **3** e1602548. https://doi.org/10.1126/sciadv.1602548

PEIXOTO, T. P. (2014). Hierarchical block structures and high-resolution model selection in large networks. *Phys. Rev. X* **4** 011047.

REYES, P. and RODRIGUEZ, A. (2016). Stochastic blockmodels for exchangeable collections of networks. arXiv preprint. Available at arXiv:1606.05277.

RIVERA-HUTINEL, A., BUSTAMANTE, R. O., MARÍN, V. H. and MEDEL, R. (2012). Effects of sampling completeness on the structure of plant–pollinator networks. *Ecology* **93** 1593–1603.

SANDER, E. L., WOOTTON, J. T. and ALLESINA, S. (2015). What can interaction webs tell us about species roles? *PLoS Comput. Biol.* **11** e1004330. https://doi.org/10.1371/journal.pcbi.1004330

SIGNORELLI, M. and WIT, E. C. (2020). Model-based clustering for populations of networks. *Stat. Model.* **20** 9–29. MR4052400 https://doi.org/10.1177/1471082X19871128

ŠKULJ, D. and ŽIBERNA, A. (2022). Stochastic blockmodeling of linked networks. *Soc. Netw.* **70** 240–252.

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742 https://doi.org/10.1007/s003579900004

STANLEY, N., SHAI, S., TAYLOR, D. and MUCHA, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **3** 95–105. MR3515211 https://doi.org/10.1109/TNSE.2016.2537545

SWEET, T. M., FLYNT, A. and CHOI, D. (2019). Clustering ensembles of social networks. *Netw. Sci.* **7** 141–159. https://doi.org/10.1017/nws.2019.2

SWEET, T. M., THOMAS, A. C. and JUNKER, B. W. (2014). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods*. CRC Press, Boca Raton, FL. MR3380043

THOMPSON, R. M. and TOWNSEND, C. R. (2003). Impacts on stream food webs of native and exotic forest: An intercontinental comparison. *Ecology* **84**. 145–161. https://doi.org/10.1890/0012-9658(2003)084[0145:IOSFWO]2.0.CO;2

VISSAULT, S., CAZELLES, K., BERGERON, G., MERCIER, B., VIOLET, C. and GRAVEL, D. (2020). rmangal: An R package to interact with Mangal database. R package version 2.0.2.

WHITE, H. C., BOORMAN, S. A. and BREIGER, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *Amer. J. Sociol.* **81** 730–780.

WILLS, P. and MEYER, F. G. (2020). Metrics for graph comparison: A practitioner's guide. *PLoS ONE* **15** e0228728. https://doi.org/10.1371/journal.pone.0228728

YIN, F., SHEN, W. and BUTTS, C. T. (2022). Finite mixtures of ERGMs for modeling ensembles of networks. *Bayesian Anal.* **17** 1153–1191. MR4506025 https://doi.org/10.1214/21-ba1298

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1236–1252
https://doi.org/10.1214/23-AOAS1832

ATHLETE RATING IN MULTICOMPETITOR GAMES WITH SCORED OUTCOMES VIA MONOTONE TRANSFORMATIONS

BY JONATHAN CHE[a] AND MARK GLICKMAN[b]

Department of Statistics, Harvard University, [a]*jche@g.harvard.edu,* [b]*glickman@fas.harvard.edu*

Sports organizations often want to estimate athlete strengths. For games with scored outcomes, a common approach is to assume observed game scores follow a normal distribution conditional on athletes' latent abilities, which may change over time. In many games, however, this assumption of conditional normality does not hold. To estimate athletes' time-varying latent abilities using nonnormal game score data, we propose a Bayesian dynamic linear model with flexible monotone response transformations. Our model learns nonlinear monotone transformations to address nonnormality in athlete scores and can be easily fit using standard regression and optimization routines, which we implement in the `dlmt` package in `R`. We demonstrate our method on data from several Olympic sports, including biathlon, diving, rugby, and fencing.

REFERENCES

ATKINSON, A. C. and SHEPHARD, N. (1996). Deletion diagnostics for transformations of time series. *J. Forecast.* **15** 1–17.

AUGER-MÉTHÉ, M., NEWMAN, K., COLE, D., EMPACHER, F., GRYBA, R., KING, A. A., LEOS-BARAJAS, , FLEMMING, J. M., NIELSEN, A. et al. (2021). A guide to state–space modeling of ecological time series. *Ecol. Monogr.* **4** 1–32.

BAKER, R. D. and MCHALE, I. G. (2015a). Deterministic evolution of strength in multiple comparisons models: Who is the greatest golfer? *Scand. J. Stat.* **42** 180–196. MR3318031 https://doi.org/10.1111/sjos.12101

BAKER, R. D. and MCHALE, I. G. (2015b). Time varying ratings in association football: The all-time greatest team is.... *J. Roy. Statist. Soc. Ser. A* **178** 481–492. MR3300014 https://doi.org/10.1111/rssa.12060

CARON, F. and TEH, Y. W. (2012). Bayesian nonparametric models for ranked data. In *Proceedings of the* 25*th International Conference on Neural Information Processing Systems* 1520–1528.

CATTELAN, M., VARIN, C. and FIRTH, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 135–150. MR3042315 https://doi.org/10.1111/j.1467-9876.2012.01046.x

CHE, J. and GLICKMAN, M. (2024). Supplement to "Athlete rating in multicompetitor games with scored outcomes via monotone transformations." https://doi.org/10.1214/23-AOAS1832SUPPA, https://doi.org/10.1214/23-AOAS1832SUPPB

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis. Texts in Statistical Science Series*. CRC Press, London. MR1385925

GLICKMAN, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 377–394.

GLICKMAN, M. E. (2001). Dynamic paired comparison models with stochastic variances. *J. Appl. Stat.* **28** 673–689. MR1862491 https://doi.org/10.1080/02664760120059219

GLICKMAN, M. E. and HENNESSY, J. (2015). A stochastic rank ordered logit model for rating multi-competitor games and sports. *J. Quant. Anal. Sports* **11** 131–144.

GLICKMAN, M. E. and STERN, H. S. (1998). A state-space model for National Football League scores. *J. Amer. Statist. Assoc.* **93** 25–35.

HARVILLE, D. (1977). The use of linear-model methodology to rate high school or college football teams. *J. Amer. Statist. Assoc.* **72** 278–289.

HARVILLE, D. A. (2003). The selection or seeding of college basketball or football teams for postseason competition. *J. Amer. Statist. Assoc.* **98** 17–27. MR1977197 https://doi.org/10.1198/016214503388619058

HERBRICH, R., MINKA, T. and GRAEPEL, T. (2006). Trueskill™: A Bayesian skill rating system. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems* 569–576.

Key words and phrases. Dynamic linear model, Kalman filter, monotone spline.

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

HOTZ-BEHOFSITS, C., HUBER, F. and ZÖRNER, T. O. (2018). Predicting crypto-currencies using sparse non-Gaussian state space models. *J. Forecast.* **37** 627–640. MR3850512 https://doi.org/10.1002/for.2524

INGRAM, M. (2019). A point-based Bayesian hierarchical model to predict the outcome of tennis matches. *J. Quant. Anal. Sports* **15** 313–325.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning*: *With Applications in R*. *Springer Texts in Statistics* **103**. Springer, New York. MR3100153 https://doi.org/10.1007/978-1-4614-7138-7

KING, B. and KOWAL, D. R. (2021). Warped dynamic linear models for time series of counts. arXiv preprint. Available at arXiv:2110.14790.

KOVALCHIK, S. (2020). Extension of the Elo rating system to margin of victory. *Int. J. Forecast.* **36** 1329–1341.

KOWAL, D. R. and CANALE, A. (2020). Simultaneous transformation and rounding (STAR) models for integer-valued data. *Electron. J. Stat.* **14** 1744–1772. MR4083734 https://doi.org/10.1214/20-EJS1707

LENK, P. J. and TSAI, C.-L. (1990). Transformations and dynamic linear models. *J. Forecast.* **9** 219–232.

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

LOPEZ, M. J., MATTHEWS, G. J. and BAUMER, B. S. (2018). How often does the best team win? A unified approach to understanding randomness in North American sport. *Ann. Appl. Stat.* **12** 2483–2516. MR3875709 https://doi.org/10.1214/18-AOAS1165

MCKEOUGH, K. (2020). A tale of two multi-phase inference applications. Ph.D. thesis, Harvard Univ. MR4272268

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

PLACKETT, R. L. (1975). The analysis of permutations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **24** 193–202. MR0391338 https://doi.org/10.2307/2346567

PRADO, R. and WEST, M. (2010). *Time Series*: *Modeling*, *Computation*, *and Inference*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2655202

RAMSAY, J. O. (1988). Monotone regression splines in action. *Statist. Sci.* **3** 425–441. https://doi.org/10.1214/SS/1177012761

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

STAN DEVELOPMENT TEAM (2021). RStan: The R interface to Stan. R package version 2.21.3.

VARADHAN, R. (2015). alabama: Constrained nonlinear optimization. R package version 2015.3-1.

WAN, E. A. and VAN DER MERWE, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Proceedings of the IEEE* 2000 *Adaptive Systems for Signal Processing*, *Communications*, *and Control Symposium* 153–158. IEEE Press, New York. (Cat. No. 00EX373).

WANG, H., ZHANG, Y.-M., MAO, J.-X., WAN, H.-P., TAO, T.-Y. and ZHU, Q.-X. (2019). Modeling and forecasting of temperature-induced strain of a long-span bridge using an improved Bayesian dynamic linear model. *Eng. Struct.* **192** 220–232.

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80** 73–83. With discussion. MR0786598

XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2000). On the estimation of an instantaneous transformation for time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 383–397. MR1749546 https://doi.org/10.1111/1467-9868.00238

ZAPPALÀ, C., PLUCHINO, A., RAPISARDA, A., BIONDO, A. E. and SOBKOWICZ, P. (2022). On the role of chance in fencing tournaments: An agent-based approach. *PLoS ONE* **17** e0267541.

ZHOU, T. and JI, Y. (2020). Semiparametric Bayesian inference for the transmission dynamics of Covid-19 with a state-space model. *Contemp. Clin. Trials* **97** 106146. https://doi.org/10.1016/j.cct.2020.106146

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1253–1274
https://doi.org/10.1214/23-AOAS1833

ESTIMATING THE LIKELIHOOD OF ARREST FROM POLICE RECORDS IN PRESENCE OF UNREPORTED CRIMES

BY RICCARDO FOGLIATO[1,a], ARUN KUMAR KUCHIBHOTLA[2,b], ZACHARY LIPTON[3,c], DANIEL NAGIN[4,d], ALICE XIANG[5,f] AND ALEXANDRA CHOULDECHOVA[4,e]

[1]*Amazon Web Services,* [a]*riccardofogliato@gmail.com*
[2]*Department of Statistics and Data Science, Carnegie Mellon University,* [b]*arunku@stat.cmu.edu*
[3]*Department of Machine Learning, Carnegie Mellon University,* [c]*zlipton@cmu.edu*
[4]*Heinz College, Carnegie Mellon University,* [d]*dn03@andrew.cmu.edu,* [e]*achoulde@andrew.cmu.edu*
[5]*Sony AI,* [f]*Alice.Xiang@sony.com*

Many important policy decisions concerning policing hinge on our understanding of how likely various criminal offenses are to result in arrests. Since many crimes are never reported to law enforcement, estimates based on police records alone must be adjusted to account for the likelihood that each crime would have been reported to the police. In this paper we present a methodological framework for estimating the likelihood of arrest from police data that incorporates estimates of crime reporting rates computed from a victimization survey. We propose a parametric regression-based two-step estimator that: (i) estimates the likelihood of crime reporting using logistic regression with survey weights and then (ii) applies a second regression step to model the likelihood of arrest. Our empirical analysis focuses on racial disparities in arrests for violent crimes (sex offenses, robbery, aggravated and simple assaults) from 2006–2015 police records from the National Incident Based Reporting System (NIBRS), with estimates of crime reporting obtained using 2003–2020 data from the National Crime Victimization Survey (NCVS). We find that, after adjusting for unreported crimes, the likelihood of arrest computed from police records decreases significantly. We also find that, while incidents with white offenders, on average, result in arrests more often than those with black offenders, the disparities tend to be small after accounting for crime characteristics and unreported crimes.

REFERENCES

ANDREWS, D. W. K. and MONAHAN, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* **60** 953–966. MR1168742 https://doi.org/10.2307/2951574

AVAKAME, E. F., FYFE, J. J. and MCCOY, C. (1999). "Did you call the police? What did they do?" An empirical assessment of Black's theory of mobilization of law. *Justice Q.* **16** 765–792.

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

BACHMAN, R. (1998). The factors related to rape reporting behavior and arrest: New evidence from the National Crime Victimization Survey. *Crim. Justice Behav.* **25** 8–29.

BARNETT-RYAN, C., LANGTON, L. and PLANTY, M. (2014). The nation's two crime measures, 2014. US Department of Justice, Washington, DC.

BASU, D. (2011). An essay on the logical foundations of survey sampling, part one [reprint of MR0423625]. In *Selected Works of Debabrata Basu. Sel. Works Probab. Stat.* 167–206. Springer, New York. MR2807264 https://doi.org/10.1007/978-1-4419-5825-9_24

BAUMER, E. P. (2002). Neighborhood disadvantage and police notification by victims of violence. *Criminology* **40** 579–616.

BAUMER, E. P. and LAURITSEN, J. L. (2010). Reporting crime to the police, 1973–2005: A multivariate analysis of long-term trends in the National Crime Survey (NCS) and National Crime Victimization Survey (NCVS). *Criminology* **48** 131–185.

Key words and phrases. Likelihood of arrest, unreported crime, racial disparities, NIBRS, NCVS.

BECK, A. J. and BLUMSTEIN, A. (2018). Racial disproportionality in US state prisons: Accounting for the effects of racial and ethnic differences in criminal involvement, arrests, sentencing, and time served. *J. Quant. Criminol.* **34** 853–883.

BERK, R., BUJA, A., BROWN, L., GEORGE, E., KUCHIBHOTLA, A. K., SU, W. and ZHAO, L. (2021). Assumption lean regression. *Amer. Statist.* **75** 76–84. MR4203483 https://doi.org/10.1080/00031305.2019.1592781

BLUMSTEIN, A. and COHEN, J. (1979). Estimation of individual crime rates from arrest records. *J. Crim. Law Criminol.* **70** 561.

BLUMSTEIN, A. and COHEN, J. (1987). Characterizing criminal careers. *Science* **237** 985–991. https://doi.org/10.1126/science.237.4818.985

BLUMSTEIN, A., COHEN, J., PIQUERO, A. R. and VISHER, C. A. (2010). Linking the crime and arrest processes to measure variations in individual arrest risk per crime (Q). *J. Quant. Criminol.* **26** 533–548.

BLUMSTEIN, A. et al. (1986). *Criminal Careers and "Career Criminals,"* **2**. National Academies.

BÖHNING, D. and VAN DER HEIJDEN, P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Ann. Appl. Stat.* **3** 595–610. MR2750674 https://doi.org/10.1214/08-AOAS214

BRAME, R., FAGAN, J., PIQUERO, A. R., SCHUBERT, C. A. and STEINBERG, L. (2004). Criminal careers of serious delinquents in two cities. *Youth Violence Juvenile Justice* **2** 256–272.

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

BUIL-GIL, D., MEDINA, J. and SHLOMO, N. (2021). Measuring the dark figure of crime in geographic areas: Small area estimation from the crime survey for England and Wales. *Br. J. Criminol.* **61** 364–388.

BUIL-GIL, D., MORETTI, A. and LANGTON, S. H. (2021). The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology* 1–27.

BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2019a). Models as approximations I: Consequences illustrated with linear regression. *Statist. Sci.* **34** 523–544. MR4048582 https://doi.org/10.1214/18-STS693

BUJA, A., BROWN, L., KUCHIBHOTLA, A. K., BERK, R., GEORGE, E. and ZHAO, L. (2019b). Models as approximations II: A model-free theory of parametric regression. *Statist. Sci.* **34** 545–565. MR4048583 https://doi.org/10.1214/18-STS694

BYRD, J. and LIPTON, Z. (2019). What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning* 872–881. PMLR.

CERNAT, A., BUIL-GIL, D., PINA-SÁNCHEZ, J., MURRIÀ-SANGENÍS, M. et al. (2021). Estimating crime in place: Moving beyond residence location.

D'ALESSIO, S. J. and STOLZENBERG, L. (2003). Race and the probability of arrest. *Soc. Forces* **81** 1381–1397.

DUGAN, L. (2003). Domestic violence legislation: Exploring its impact on the likelihood of domestic violence, police involvement, and arrest. *Criminol. Public Policy* **2** 283–312.

FISHER, B. S., DAIGLE, L. E., CULLEN, F. T. and TURNER, M. G. (2003). Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Crim. Justice Behav.* **30** 6–38.

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. and MOLENBERGHS, G., eds. (2009) *Longitudinal Data Analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR1500110

FITZMAURICE, G. M., LAIRD, N. M. and ROTNITZKY, A. G. (1993). Regression models for discrete longitudinal responses. *Statist. Sci.* **8** 284–309. MR1243595

FOGLIATO, R., XIANG, A., LIPTON, Z., NAGIN, D. and CHOULDECHOVA, A. (2021). On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the* 2021 *AAAI/ACM Conference on AI, Ethics, and Society. AIES'*21 100–111. Assoc. Comput. Mach., New York, NY, USA. https://doi.org/10.1145/3461702.3462538

FOGLIATO, R., KUCHIBHOTLA, A. K, LIPTON, Z., NAGIN, D., XIANG, A. and CHOULDECHOVA, A. (2024). Supplement to "Estimating the likelihood of arrest from police records in presence of unreported crimes." https://doi.org/10.1214/23-AOAS1833SUPPA, https://doi.org/10.1214/23-AOAS1833SUPPB

GRAHAM, J. W., OLCHOWSKI, A. E. and GILREATH, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **8** 206–213. https://doi.org/10.1007/s11121-007-0070-9

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161. MR0518832 https://doi.org/10.2307/1912352

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

HUBBARD, A. E., AHERN, J., FLEISCHER, N. L., VAN DER LAAN, M., SATARIANO, S. A., JEWELL, N., BRUCKNER, T. and SATARIANO, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 467–474.

HUGGINS, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76** 133–140. MR0991431 https://doi.org/10.1093/biomet/76.1.133

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

KOCHEL, T. R., WILSON, D. B. and MASTROFSKI, S. D. (2011). Effect of suspect race on officers' arrest decisions. *Criminology* **49** 473–512.

LANTZ, B. and WENGER, M. R. (2019). The co-offender as counterfactual: A quasi-experimental within-partnership approach to the examination of the relationship between race and arrest. *Journal of Experimental Criminology* 1–24.

LEE, S.-M. and CHAO, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50** 88–97.

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. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LOEFFLER, C. E., HYATT, J. and RIDGEWAY, G. (2019). Measuring self-reported wrongful convictions among prisoners. *J. Quant. Criminol.* **35** 259–286.

LOHR, S. L. (2007). Comment: Struggles with survey weighting and regression modeling [MR2408951]. *Statist. Sci.* **22** 175–178. MR2408955 https://doi.org/10.1214/088342307000000159

LUMLEY, T. and SCOTT, A. (2017). Fitting regression models to survey data. *Statist. Sci.* **32** 265–278. MR3648959 https://doi.org/10.1214/16-STS605

LYTLE, D. J. (2014). The effects of suspect characteristics on arrest: A meta-analysis. *J. Crim. Justice* **42** 589–597.

MORGAN, R. E., BUREAU OF JUSTICE STATISTICS (BJS) and US DEPT OF JUSTICE AND OFFICE OF JUSTICE PROGRAMS AND UNITED STATES OF AMERICA (2017). Race and Hispanic origin of victims and offenders, 2012–2015. *Victims and Offenders* **2012** 15.

MORGAN, R. E. and TRUMAN, J. (2021). *Criminal Victimization*, 2020 **4**. National Crime Victimization Survey, Bureau of Justice Statistics, Washington, DC.

NAGIN, D. S. (2013). Deterrence in the twenty-first century. *Crime and Justice* **42** 199–263.

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

PETERSEN, C. G. J. (1896). The yearly immigration of young plaice in the Limfjord from the German sea. *Rept. Danish Biol. Sta.* **6** 1–48.

PIQUERO, A. R. and BRAME, R. W. (2008). Assessing the race–crime and ethnicity–crime relationship in a sample of serious adolescent delinquents. *Crime & Delinquency* **54** 390–422.

POLLEY, E. C. and VAN DER LAAN, M. J. (2010). Super learner in prediction.

POPE, C. E. and SNYDER, H. N. (2003). Race as a factor in juvenile arrests. Citeseer.

RACINE, J. and LI, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *J. Econometrics* **119** 99–130. MR2041894 https://doi.org/10.1016/S0304-4076(03)00157-X

RENNISON, C. M. (2010). An investigation of reporting violence to the police: A focus on Hispanic victims. *Journal of Criminal Justice* **38** 390–399.

RICHARDSON, R., SCHULTZ, J. and CRAWFORD, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, Forthcoming.

ROBERTS, A. and LYONS, C. J. (2009). Victim-offender racial dyads and clearance of lethal and nonlethal assault. *Journal of Research in Crime and Delinquency* **46** 301–326.

ROBERTS, A. and LYONS, C. J. (2011). Hispanic victims and homicide clearance by arrest. *Homicide Studies* **15** 48–73.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling. Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

SKOGAN, W. G. (1974). The validity of official crime statistics: An empirical investigation. *Social Science Quarterly* 25–38.

SKOGAN, W. G. (1977). Dimensions of the dark figure of unreported crime. *Crime & Delinquency* **23** 41–50.

STEFFENSMEIER, D., FELDMEYER, B., HARRIS, C. T. and ULMER, J. T. (2011). Reassessing trends in black violent crime, 1980–2008: Sorting out the "Hispanic effect" in uniform crime reports arrests, national crime victimization survey offender estimates, and US prisoner counts. *Criminology* **49** 197–251.

SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* **8**.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
UNITED STATES DEPARTMENT OF JUSTICE, BUREAU OF JUSTICE STATISTICS (2017). National Crime Victimization Survey, 2016. Technical Documentation.
UNITED STATES DEPARTMENT OF JUSTICE, BUREAU OF JUSTICE STATISTICS (2021). National crime victimization survey, concatenated file, [United States], 1992–2020. https://doi.org/10.3886/ICPSR38136.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2008a). National incident-based reporting system, 2006. https://doi.org/10.3886/ICPSR22407.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2008b). Uniform crime reporting program data [United States]: Police employee (LEOKA) data, 2006. https://doi.org/10.3886/ICPSR22402.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2009a). National incident-based reporting system, 2007. https://doi.org/10.3886/ICPSR25113.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2009b). Uniform crime reporting program data [United States]: Police employee (LEOKA) data, 2007. https://doi.org/10.3886/ICPSR25104.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2010a). National incident-based reporting system, 2008. https://doi.org/10.3886/ICPSR27647.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2010b). Uniform crime reporting program data [United States]: Police employee (LEOKA) data, 2008. https://doi.org/10.3886/ICPSR27646.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2011a). Uniform crime reporting: National incident-based reporting system, 2009. https://doi.org/10.3886/ICPSR30770.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2011b). Uniform crime reporting program data [United States]: Police employee (LEOKA) data, 2009. https://doi.org/10.3886/ICPSR30765.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2012a). Uniform crime reporting: National incident-based reporting system, 2010. https://doi.org/10.3886/ICPSR33530.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2012b). Uniform crime reporting program data: Police employee (LEOKA) data, 2010. https://doi.org/10.3886/ICPSR33525.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2013a). Uniform crime reporting program data: National incident-based reporting system, 2011. https://doi.org/10.3886/ICPSR34585.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2013b). Uniform crime reporting program data: Police employee (LEOKA) data, 2011. https://doi.org/10.3886/ICPSR34584.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2014a). Uniform crime reporting program data: National incident-based reporting system, 2012. https://doi.org/10.3886/ICPSR35035.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2014b). Uniform crime reporting program data: Police employee (LEOKA) data, 2012. https://doi.org/10.3886/ICPSR35020.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2015a). Uniform crime reporting program data: National incident-based reporting system, 2013. https://doi.org/10.3886/ICPSR36120.v2
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2015b). Uniform crime reporting program data: Police employee (LEOKA) data, 2013. https://doi.org/10.3886/ICPSR36119.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2016a). Uniform crime reporting program data: National incident-based reporting system, 2014. https://doi.org/10.3886/ICPSR36398.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2016b). Uniform crime reporting program data: Police employee (LEOKA) data, 2014. https://doi.org/10.3886/ICPSR36395.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2017a). Uniform crime reporting program data: National incident-based reporting system, 2015. https://doi.org/10.3886/ICPSR36795.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (2017b). Uniform crime reporting program data: Police employee (LEOKA) data, 2015. https://doi.org/10.3886/ICPSR36791.v1
UNITED STATES DEPARTMENT OF JUSTICE, FEDERAL BUREAU OF INVESTIGATION (FBI) (2019). 2019 National Incident-Based Reporting System User Manual.
VAN DER HEIJDEN, P. G. M., BUSTAMI, R., CRUYFF, M. J. L. F., ENGBERSEN, G. and VAN HOUWELINGEN, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Stat. Model.* **3** 305–322. MR2012155 https://doi.org/10.1191/1471082X03st057oa

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

WHITE, H. (2014). *Asymptotic Theory for Econometricians*. Academic Press, San Diego.

XIE, M. and BAUMER, E. P. (2019a). Neighborhood immigrant concentration and violent crime reporting to the police: A multilevel analysis of data from the National Crime Victimization Survey. *Criminology* **57** 237–267.

XIE, M. and BAUMER, E. P. (2019b). Crime victims' decisions to call the police: Past research and new directions. *Annual Review of Criminology*.

XIE, M. and LAURITSEN, J. L. (2012). Racial context and crime reporting: A test of Black's stratification hypothesis. *Journal of Quantitative Criminology* **28** 265–293.

XIE, M. and LYNCH, J. P. (2017). The effects of arrest, reporting to the police, and victim services on intimate partner violence. *Journal of Research in Crime and Delinquency* **54** 338–378.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1275–1293
https://doi.org/10.1214/23-AOAS1834

SEMIPARAMETRIC BIVARIATE HIERARCHICAL STATE SPACE MODEL WITH APPLICATION TO HORMONE CIRCADIAN RELATIONSHIP

BY MENGYING YOU[a] AND WENSHENG GUO[b]

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,
[a]*mengying.you@pennmedicine.upenn.edu,* [b]*wguo@upenn.edu*

The adrenocorticotropic hormone and cortisol play critical roles in stress regulation and the sleep-wake cycle. Most research has been focused on how the two hormones regulate each other in terms of short-term pulses. Few studies have been conducted on the circadian relationship between the two hormones and how it differs between normal and abnormal groups. The circadian patterns are difficult to model as parametric functions. Directly extending univariate functional mixed effects models would result in a large dimensional problem and a challenging nonparametric inference. In this article we propose a semiparametric bivariate hierarchical state space model in which each hormone profile is modeled by a hierarchical state space model with nonparametric population-average and subject-specific components. The bivariate relationship is constructed by concatenating two latent independent subject-specific random functions specified by a design matrix, leading to a parametric inference on the correlation. We propose a computationally efficient state-space EM algorithm for estimation and inference. We apply the proposed method to a study of chronic fatigue syndrome and fibromyalgia and discover an erratic regulation pattern in the patient group in contrast to a circadian regulation pattern conforming to the day–night cycle in the control group.

REFERENCES

ANSLEY, C. F., KOHN, R. and WONG, C.-M. (1993). Nonparametric spline regression with prior information. *Biometrika* **80** 75–88. MR1225215 https://doi.org/10.1093/biomet/80.1.75

ANTONIADIS, A. and SAPATINAS, T. (2007). Estimation and inference in functional mixed-effects models. *Comput. Statist. Data Anal.* **51** 4793–4813. MR2364541 https://doi.org/10.1016/j.csda.2006.09.038

BOONEN, E., MEERSSEMAN, P., VERVENNE, H., MEYFROIDT, G., GUÏZA, F. and WOUTERS, P. J. (2014). Reduced nocturnal ACTH-driven cortisol secretion during critical illness. *Amer. J. Physiol: Endocrinol. Metab.* **306** E883–E892.

CARROLL, C., MÜLLER, H.-G. and KNEIP, A. (2021). Cross-component registration for multivariate functional data, with application to growth curves. *Biometrics* **77** 839–851. MR4320661 https://doi.org/10.1111/biom.13340

CHIOU, J.-M., CHEN, Y.-T. and YANG, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statist. Sinica* **24** 1571–1596. MR3308652

CHIOU, J.-M. and MÜLLER, H.-G. (2016). A pairwise interaction model for multivariate functional and longitudinal data. *Biometrika* **103** 377–396. MR3509893 https://doi.org/10.1093/biomet/asw007

CRAINICEANU, C. M., STAICU, A.-M., RAY, S. and PUNJABI, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Stat. Med.* **31** 3223–3240. MR2993623 https://doi.org/10.1002/sim.5439

CROFFORD, L. J. (2015). Chronic pain: Where the body meets the brain. *Trans. Amer. Clin. Climatol. Assoc.* **126** 167–183.

CROFFORD, L. J., YOUNG, E. A., ENGLEBERG, N. C., KORSZUN, A., BRUCKSCH, C. B., MCCLURE, L. A., BROWN, M. B. and DEMITRACK, M. A. (2004). Basal circadian and pulsatile ACTH and cortisol secretion in patients with fibromyalgia and/or chronic fatigue syndrome. *Brain Behav. Immun.* **18** 314–325. https://doi.org/10.1016/j.bbi.2003.12.011

Key words and phrases. Circadian rhythm, functional mixed effects model, state space EM algorithm, time-varying correlation.

CZEISLER, C. A. and WATERHOUSE, J. M. (1995). The effect of light on the human circadian pacemaker. *Circadian Clocks and Their Adjust*. **183** 254–290.
DALLMAN, M. F., STRACK, A. M., AKANA, S. F., BRADBURY, M. J., HANSON, E. S., SCRIBNER, K. A. and SMITH, M. (1993). Feast and famine: Critical role of glucocorticoids with insulin in daily energy flow. *Front. Neuroendocrinol*. **14** 303–347.
DE JONG, P. (1989). Smoothing and interpolation with the state-space model. *J. Amer. Statist. Assoc*. **84** 1085–1088. MR1134497
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
DUBIN, J. A. and MÜLLER, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *J. Amer. Statist. Assoc*. **100** 872–881. MR2201015 https://doi.org/10.1198/016214504000001989
DUMBELL, R., MATVEEVA, O. and OSTER, H. (2016). Circadian clocks, stress, and immunity. *Front. Endocrinol*. **7** 37. https://doi.org/10.3389/fendo.2016.00037
ENGLER, D., PHAM, T., LIU, J.-P., FULLERTON, M. J., CLARKE, I. J. and FUNDER, J. W. (1990). Studies of the regulation of the hypothalamic-pituitary-adrenal axis in sheep with hypothalamic-pituitary disconnection. II. Evidence for in vivo ultradian hypersecretion of proopiomelanocortin peptides by the isolated anterior and intermediate pituitary. *Endocrinology* **127** 1956–1966.
FOCKE, C. M. B. and IREMONGER, K. J. (2020). Rhythmicity matters: Circadian and ultradian patterns of HPA axis activity. *Mol. Cell. Endocrinol*. **501** 110652.
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
GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869 https://doi.org/10.1007/978-1-4614-5369-7
GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x
HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc*. **113** 649–659. MR3832216 https://doi.org/10.1080/01621459.2016.1273115
HE, G., MÜLLER, H.-G. and WANG, J.-L. (2003). Functional canonical analysis for square integrable stochastic processes. *J. Multivariate Anal*. **85** 54–77. MR1978177 https://doi.org/10.1016/S0047-259X(02)00056-8
HE, J. (2014). Functional correlations to quantify functional connectivity in brain imaging. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Univ. California, Davis. MR3337706
JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. MR1789811 https://doi.org/10.1093/biomet/87.3.587
KOOPMAN, S. J. and DURBIN, J. (2000). Fast filtering and smoothing for multivariate state space models. *J. Time Series Anal*. **21** 281–296. MR1766960 https://doi.org/10.1111/1467-9892.00186
LAI, T., ZHANG, Z., WANG, Y. and KONG, L. (2021). Testing independence of functional variables by angle covariance. *J. Multivariate Anal*. **182** Paper No. 104711. MR4187269 https://doi.org/10.1016/j.jmva.2020.104711
LEURGANS, S. E., MOYEED, R. A. and SILVERMAN, B. W. (1993). Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55** 725–740. MR1223939
LI, C., XIAO, L. and LUO, S. (2020). Fast covariance estimation for multivariate sparse functional data. *Stat* **9** e245. MR4116315
LIGHTMAN, S. L., BIRNIE, M. T. and CONWAY-CAMPBELL, B. L. (2020). Dynamics of ACTH and cortisol secretion and implications for disease. *Endocr. Rev.* **41** 470–490.
LIU, Z., CAPPOLA, A. R., CROFFORD, L. J. and GUO, W. (2014). Modeling bivariate longitudinal hormone profiles by hierarchical state space models. *J. Amer. Statist. Assoc*. **109** 108–118. MR3180550 https://doi.org/10.1080/01621459.2013.830071
LIU, Z. and GUO, W. (2012). Functional mixed effects models. *Wiley Interdiscip. Rev.: Comput. Stat*. **4** 527–534.
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
QIN, L. (2004). Functional models using smoothing splines, a state space approach. Dissertation, Univ. Pennsylvania.
QIN, L. and GUO, W. (2006). Functional mixed-effects model for periodic data. *Biostatistics* **7** 225–234.
RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314 https://doi.org/10.1111/j.0006-341X.2001.00253.x
SANG, P., WANG, L. and CAO, J. (2019). Weighted empirical likelihood inference for dynamical correlations. *Comput. Statist. Data Anal*. **131** 194–206. MR3906804 https://doi.org/10.1016/j.csda.2018.07.003

SHUMWAY, R. H. and STOFFER, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*. **3** 253–264. https://doi.org/10.1111/j.1467-9892.1982.tb00349.x

VELDHUIS, J. D., IRANMANESH, A., NAFTOLOWITZ, D., TATHAM, N., CASSIDY, F. and CARROLL, B. J. (2001). Corticotropin secretory dynamics in humans under low glucocorticoid feedback. *J. Clin. Endocrinol. Metab.* **86** 5554–5563.

VOLKMANN, A., STÖCKER, A., SCHEIPL, F. and GREVEN, S. (2021). Multivariate functional additive mixed models. *Stat. Model.* **23** 303–326. MR4624326 https://doi.org/10.1177/1471082X211056158

YOU, M. and GUO, W. (2024). Supplement to "Semiparametric bivariate hierarchical state space model with application to hormone circadian relationship." https://doi.org/10.1214/23-AOAS1834SUPP

YOUNG, E. A., CARLSON, N. E. and BROWN, M. B. (2001). Twenty-four-hour ACTH and cortisol pulsatility in depressed women. *Neuropsychopharmacology* **25** 267–276.

ZHOU, L., HUANG, J. Z. and CARROLL, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika* **95** 601–619. MR2443178 https://doi.org/10.1093/biomet/asn035

ZHOU, Y., LIN, S.-C. and WANG, J.-L. (2018). Local and global temporal correlations for longitudinal data. *J. Multivariate Anal.* **167** 1–14. MR3830630 https://doi.org/10.1016/j.jmva.2018.03.015

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1294–1318
https://doi.org/10.1214/23-AOAS1835

TENSOR QUANTILE REGRESSION WITH LOW-RANK TENSOR TRAIN ESTIMATION

BY ZIHUAN LIU[1,a], CHEUK YIN LEE[2,c] AND HEPING ZHANG[1,b]

[1]*Department of Biostatistics, Yale University,* [a]*liuzihua@msu.edu,* [b]*heping.zhang@yale.edu*
[2]*School of Science and Engineering, Chinese University of Hong Kong,* [c]*leecheukyin@cuhk.edu.cn*

Neuroimaging studies often involve predicting a scalar outcome from an array of images collectively called tensor. The use of magnetic resonance imaging (MRI) provides a unique opportunity to investigate the structures of the brain. To learn the association between MRI images and human intelligence, we formulate a scalar-on-image quantile regression framework. However, the high dimensionality of the tensor makes estimating the coefficients for all elements computationally challenging. To address this, we propose a low-rank coefficient array estimation algorithm, based on tensor train (TT) decomposition, which we demonstrate can effectively reduce the dimensionality of the coefficient tensor to a feasible level while ensuring adequacy to the data. Our method is more stable and efficient compared to the commonly used canonic polyadic rank approximation-based method. We also propose a generalized lasso penalty on the coefficient tensor to take advantage of the spatial structure of the tensor, further reduce the dimensionality of the coefficient tensor, and improve the interpretability of the model. The consistency and asymptotic normality of the TT estimator are established under some mild conditions on the covariates and random errors in quantile regression models. The rate of convergence is obtained with regularization under the total variation penalty. Extensive numerical studies, including both synthetic and real MRI imaging data, are conducted to examine the empirical performance of the proposed method and its competitors.

REFERENCES

AHMED, T., RAJA, H. and BAJWA, W. U. (2020). Tensor regression using low-rank and sparse Tucker decompositions. *SIAM J. Math. Data Sci.* **2** 944–966. MR4161310 https://doi.org/10.1137/19M1299335

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841 https://doi.org/10.1214/10-AOS827

BILKER, W. B., HANSEN, J. A., BRENSINGER, C. M., RICHARD, J., GUR, R. E. and GUR, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment* **19** 354–369.

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

BRANTLEY, H. L., GUINNESS, J. and CHI, E. C. (2020). Baseline drift estimation for air quality data using quantile trend filtering. *Ann. Appl. Stat.* **14** 585–604. MR4117821 https://doi.org/10.1214/19-AOAS1318

CHAN, E., MACPHERSON, S. E., BOZZALI, M., SHALLICE, T. and CIPOLOTTI, L. (2018). The influence of fluid intelligence, executive functions and premorbid intelligence on memory in frontal patients. *Front. Psychol.* **9** 926. https://doi.org/10.3389/fpsyg.2018.00926

CHEN, P.-Y., CHEN, C.-L., HSU, Y.-C., TSENG, W.-Y. I. et al. (2020). Fluid intelligence is associated with cortical volume and white matter tract integrity within multiple-demand system across adult lifespan. *NeuroImage* **212** 116576.

CHEN, Z., BATSELIER, K., SUYKENS, J. A. K. and WONG, N. (2018). Parallelized tensor train learning of polynomial classifiers. *IEEE Trans. Neural Netw. Learn. Syst.* **29** 4621–4632. MR3875026 https://doi.org/10.1109/tnnls.2017.2771264

Key words and phrases. Tensor regression, conditional quantile, tensor train (TT) decomposition, total variation.

CICHOCKI, A., MANDIC, D., PHAN, A.-H., CAIAFA, C., ZHOU, G., ZHAO, Q. and LATHAUWER, L. (2014). Tensor decompositions for signal processing applications from two-way to multiway component analysis. *IEEE Signal Process. Mag.* **32**.

COLOM, R., KARAMA, S., JUNG, R. E. and HAIER, R. J. (2022). Human intelligence and brain networks. *Dialogues Clin. Neurosci.*.

DA SILVA, C. and HERRMANN, F. J. (2015). Optimization on the hierarchical Tucker manifold—applications to tensor completion. *Linear Algebra Appl.* **481** 131–173. MR3349649 https://doi.org/10.1016/j.laa.2015.04.015

DEARY, I. (2000). *Looking down on Human Intelligence: From Psychometrics to the Brain* **36**. OUP Oxford, Oxford.

FENG, L., BI, X. and ZHANG, H. (2021). Brain regions identified as being associated with verbal reasoning through the use of imaging regression via internal variation. *J. Amer. Statist. Assoc.* **116** 144–158. MR4227681 https://doi.org/10.1080/01621459.2020.1766468

GE, T., CHEN, C.-Y., DOYLE, A., VETTERMANN, R., TUOMINEN, L., HOLT, D., SABUNCU, M. and SMOLLER, J. (2018). The shared genetic basis of educational attainment and cerebral cortical morphology. *Cereb. Cortex* **29**.

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

GONG, Q.-Y., SLUMING, V., MAYES, A., KELLER, S., BARRICK, T., CEZAYIRLI, E. and ROBERTS, N. (2005). Voxel-based morphometry and stereology provide convergent evidence of the importance of medial prefrontal cortex for fluid intelligence in healthy adults. *NeuroImage* **25** 1175–1186.

GUO, W., KOTSIA, I. and PATRAS, I. (2012). Tensor learning for regression. *IEEE Trans. Image Process.* **21** 816–827. MR2932176 https://doi.org/10.1109/TIP.2011.2165291

HAIER, R. J., JUNG, R. E., YEO, R. A., HEAD, K. and ALKIRE, M. T. (2004). Structural brain variation and general intelligence. *NeuroImage* **23** 425–433.

HE, X., NG, P. and PORTNOY, S. (1998). Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 537–550. MR1625950 https://doi.org/10.1111/1467-9868.00138

HITCHCOCK, F. (1927). The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.* **6**.

JU, F., SUN, Y., GAO, J., HU, Y. and YIN, B. (2018). Vectorial dimension reduction for tensors based on Bayesian inference. *IEEE Trans. Neural Netw. Learn. Syst.* **29** 4579–4592. MR3875023 https://doi.org/10.1109/tnnls.2017.2739131

KANG, H., OMBAO, H., LINKLETTER, C., LONG, N. and BADRE, D. (2012). Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. *J. Amer. Statist. Assoc.* **107** 568–577. MR2980068 https://doi.org/10.1080/01621459.2012.664503

KE, B., ZHAO, W. and WANG, L. (2023). Smoothed tensor quantile regression estimation for longitudinal data. *Comput. Statist. Data Anal.* **178** 107609. MR4488685 https://doi.org/10.1016/j.csda.2022.107609

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657 https://doi.org/10.1017/CBO9780511754098

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417 https://doi.org/10.1093/biomet/81.4.673

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** 95–138. MR0444690 https://doi.org/10.1016/0024-3795(77)90069-6

LANGESLAG, S. J., SCHMIDT, M., GHASSABIAN, A., JADDOE, V. W., HOFMAN, A., VAN DER LUGT, A., VERHULST, F. C., TIEMEIER, H. and WHITE, T. J. (2013). Functional connectivity between parietal and frontal brain regions and intelligence in young children: The generation R study. *Hum. Brain Mapp.* **34** 3299–3307.

LI, C. and ZHANG, H. (2021). Tensor quantile regression with application to association between neuroimages and human intelligence. *Ann. Appl. Stat.* **15** 1455–1477. MR4316657 https://doi.org/10.1214/21-aoas1475

LI, P., SOFUOGLU, S. E., AVIYENTE, S. and MAITI, T. (2022). Coupled support tensor machine classification for multimodal neuroimaging data. *Stat. Anal. Data Min.* **15** 797–818. MR4524850 https://doi.org/10.1002/sam.11587

LI, X., ZHOU, H. and LI, L. (2013). Tucker tensor regression and neuroimaging analysis. *Stat. Biosci.* **10**.

LIANG, J., HÄRDLE, W. K. and TIAN, M. (2023). Imputed quantile tensor regression for near-sited spatial-temporal data. *Comput. Statist. Data Anal.* **182** 107713. MR4550796 https://doi.org/10.1016/j.csda.2023.107713

LIU, Y., LIU, J. and ZHU, C. (2020). Low-rank tensor train coefficient array estimation for tensor-on-tensor regression. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 5402–5411. MR4189257

LIU, Z., LEE, C. Y. and ZHANG, H. (2024). Supplement to "Tensor quantile regression with low-rank tensor train estimation." https://doi.org/10.1214/23-AOAS1835SUPPA, https://doi.org/10.1214/23-AOAS1835SUPPB

LU, W., ZHU, Z. and LIAN, H. (2020). High-dimensional quantile tensor regression. *J. Mach. Learn. Res.* **21** 250. MR4213431

LUDERS, E., GASER, C., JANCKE, L. and SCHLAUG, G. (2004). A voxel-based approach to gray matter asymmetries. *NeuroImage* **22** 656–664. https://doi.org/10.1016/j.neuroimage.2004.01.032

MEYER, V. (1959). Cognitive changes following temporal lobectomy for relief of temporal lobe epilepsy. *A.M.A. Arch. Neurol. Psych.* **81** 299–309.

OSELEDETS, I. V. (2011). Tensor-train decomposition. *SIAM J. Sci. Comput.* **33** 2295–2317. MR2837533 https://doi.org/10.1137/090752286

QIU, M.-G., YE, Z., LI, Q.-Y., LIU, G.-J., XIE, B. and WANG, J. (2011). Changes of brain structure and function in ADHD children. *Brain Topogr.* **24** 243–252.

RHEIN, C., MÜHLE, C., RICHTER-SCHMIDINGER, T., ALEXOPOULOS, P., DOERFLER, A. and KORNHUBER, J. (2014). Neuroanatomical correlates of intelligence in healthy young adults: The role of basal ganglia volume. *PLoS ONE* **9** e93623. https://doi.org/10.1371/journal.pone.0093623

ROHWEDDER, T. and USCHMAJEW, A. (2013). On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM J. Numer. Anal.* **51** 1134–1162. MR3038114 https://doi.org/10.1137/110857520

SHAW, P., GREENSTEIN, D., LERCH, J., CLASEN, L., LENROOT, R., GOGTAY, N., EVANS, A., RAPOPORT, J. and GIEDD, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature* **440** 676–679.

SI, Y., ZHANG, Y. and LI, G. (2022). An efficient tensor regression for high-dimensional data.

SIDIROPOULOS, N. D., DE LATHAUWER, L., FU, X., HUANG, K., PAPALEXAKIS, E. E. and FALOUTSOS, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* **65** 3551–3582. MR3666587 https://doi.org/10.1109/TSP.2017.2690524

TOGA, A. W. and THOMPSON, P. M. (2005). Genetics of brain structure and intelligence. *Annu. Rev. Neurosci.* **28** 1–23. https://doi.org/10.1146/annurev.neuro.28.061604.135655

UĞURBIL, K., XU, J., AUERBACH, E. J., MOELLER, S., VU, A. T., DUARTE-CARVAJALINO, J. M., LENGLET, C., WU, X., SCHMITTER, S. et al. (2013). Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *NeuroImage* **80** 80–104.

USCHMAJEW, A. and VANDEREYCKEN, B. (2013). The geometry of algorithms using hierarchical tensors. *Linear Algebra Appl.* **439** 133–166. MR3045227 https://doi.org/10.1016/j.laa.2013.03.016

VAN ESSEN, D. C., SMITH, S. M., BARCH, D. M., BEHRENS, T. E., YACOUB, E., UGURBIL, K., CONSORTIUM, W.-M. H. et al. (2013). The WU-Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

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

WATKINS, K. E., PAUS, T., LERCH, J. P., ZIJDENBOS, A., COLLINS, D. L., NEELIN, P., TAYLOR, J., WORSLEY, K. J. and EVANS, A. C. (2001). Structural asymmetries in the human brain: A voxel-based statistical analysis of 142 MRI scans. *Cereb. Cortex* **11** 868–877.

WEI, B., PENG, L., GUO, Y., MANATUNGA, A. and STEVENS, J. (2023). Tensor response quantile regression with neuroimaging data. *Biometrics* **79** 1947–1958. MR4643968 https://doi.org/10.1111/biom.13809

YUANKAI, W., TAN, H., LI, Y., ZHANG, J. and CHEN, X. (2018). A fused CP factorization method for incomplete tensors. *IEEE Trans. Neural Netw. Learn. Syst.* **PP** 1–14.

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499

ZNIYED, Y., MIRON, S., BOYER, R. and BRIE, D. (2019). Uniqueness of tensor train decomposition with linear dependencies. In 2019 *IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* (*CAMSAP*) 460–464. IEEE, Los Alamitos.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1319–1336
https://doi.org/10.1214/23-AOAS1836

RISK-AWARE RESTRICTED OUTCOME LEARNING FOR INDIVIDUALIZED TREATMENT REGIMES OF SCHIZOPHRENIA

BY SHUYING ZHU[1,a], WEINING SHEN[2,b], HAODA FU[3,d] AND ANNIE QU[2,c]

[1]*Meta, Seattle,* [a]*shuyingzhu@meta.com*
[2]*Department of Statistics, University of California, Irvine,* [b]*weinings@uci.edu,* [c]*aqu2@uci.edu*
[3]*Eli Lilly and Company,* [d]*fu_haoda@lilly.com*

Schizophrenia is a severe mental disorder that distorts patients' perception of reality, and its treatment with antipsychotics can lead to significant side effects. Despite the heterogeneity in patient responses to treatments, most existing studies on individualized treatment regimes only focus on optimizing treatment efficacy, disregarding potential negative effects. To fill this gap, we propose a restricted outcome weighted learning method that optimizes efficacy outcomes while adhering to individual-level negative effect constraints. Our method is developed for multistage treatment decision problems that include single-stage decision as a special case. We propose an efficient learning algorithm that utilizes the difference-of-convex algorithm and the Lagrange multiplier to solve nonconvex optimization with nonconvex risk constraints. We also establish theoretical properties, including Fisher consistency and strong duality results, for the proposed method. We apply our method to a clinical study to design effective schizophrenia treatment [Stroup et al. (*Schizophr. Bull.* **29** (2003) 15–31)] and find that our approach reduces side-effect risk by at least 22.5% and improves efficacy by at least 26.3% compared to competing methods. In addition, we discover that certain covariates, such as the PANSS score, clinician global impressions severity score, and BMI, have a significant impact on controlling side effects and determining optimal treatment recommendations. These results are valuable in identifying subgroups of patients who need special attention when prescribing more aggressive treatment plans.

REFERENCES

ALMIRALL, D., TEN HAVE, T. and MURPHY, S. A. (2010). Structural nested mean models for assessing time-varying effect moderation. *Biometrics* **66** 131–139. MR2756699 https://doi.org/10.1111/j.1541-0420.2009.01238.x

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

BUTLER, E. L., LABER, E. B., DAVIS, S. M. and KOSOROK, M. R. (2018). Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics* **74** 18–26. MR3777922 https://doi.org/10.1111/biom.12743

CLIFTON, J. and LABER, E. (2020). *Q*-learning: Theory and applications. *Annu. Rev. Stat. Appl.* **7** 279–301. MR4104194 https://doi.org/10.1146/annurev-statistics-031219-041220

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

FANG, E. X., WANG, Z. and WANG, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *J. Amer. Statist. Assoc.* **118** 1733–1746. MR4646602 https://doi.org/10.1080/01621459.2021.2008402

FORNARO, M., ANASTASIA, A., VALCHERA, A., CARANO, A., ORSOLINI, L., VELLANTE, F., RAPINI, G., OLIVIERI, L., DI NATALE, S. et al. (2019). The FDA "black box" warning on antidepressant suicide risk in young adults: More harm than benefits? *Frontiers in Psychiatry* **10** 294.

FRIEDMAN, R. A. (2014). Antidepressants' black-box warning—10 years later. *N. Engl. J. Med.* **371** 1666–1668. https://doi.org/10.1056/NEJMp1408480

Key words and phrases. Dynamic treatment regimes, individual-level risk control, individualized treatment regimes, outcome weighted learning, restricted optimization, side effects.

GEWANDTER, J. S., MCDERMOTT, M. P., EVANS, S., KATZ, N. P., MARKMAN, J. D., SIMON, L. S., TURK, D. C. and DWORKIN, R. H. (2021). Composite outcomes for pain clinical trials: Considerations for design and interpretation. *Pain* **162** 1899–1905.

GILLMAN, M. W. and HAMMOND, R. A. (2016). Precision treatment and precision prevention: Integrating "below and above the skin". *JAMA Pediatr* **170** 9–10. https://doi.org/10.1001/jamapediatrics.2015.2786

HODSON, R. (2016). Precision medicine. *Nature* **537** S49. https://doi.org/10.1038/537S49a

HUANG, X., SHI, L. and SUYKENS, J. A. K. (2014). Ramp loss linear programming support vector machine. *J. Mach. Learn. Res.* **15** 2185–2211. MR3231595

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

LAKKARAJU, H. and RUDIN, C. (2017). Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics* 166–175. PMLR.

LAVORI, P. W. and DAWSON, R. (2004). Dynamic treatment regimes: Practical design considerations. *Clin. Trials* **1** 9–20. https://doi.org/10.1191/1740774s04cn002oa

LEE, J., THALL, P. F., JI, Y. and MÜLLER, P. (2015). Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Amer. Statist. Assoc.* **110** 711–722. MR3367259 https://doi.org/10.1080/01621459.2014.926815

LIU, L. and KENNEDY, E. H. (2021). Median optimal treatment regimes. ArXiv preprint. Available at arXiv:2103.01802.

LUCKETT, D. J., LABER, E. B., KIM, S. and KOSOROK, M. R. (2021). Estimation and optimization of composite outcomes. *J. Mach. Learn. Res.* **22** Paper No. 167. MR4318523

MCGURK, S. R., GREEN, M. F., WIRSHING, W. C., AMES, D., MARSHALL, B., MARDER, S. R. and MINTZ, J. (1997). The effects of risperidone vs haloperidol on cognitive functioning in treatment-resistant schizophrenia: The trail making test. *CNS Spectr.* **2** 60–64.

MOREAU, D. and WIEBELS, K. (2021). Assessing change in intervention research: The benefits of composite outcomes. *Adv. Methods Pract. Psychol. Sci.* **4** 1–14.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

POPLI, A. P., KONICKI, P. E., JURJUS, G. J., FULLER, M. A. and JASKIW, G. E. (1997). Clozapine and associated diabetes mellitus. *J. Clin. Psychiatry* **58** 108–111. https://doi.org/10.4088/jcp.v58n0304

POTRA, F. A. and WRIGHT, S. J. (2000). Interior-point methods. *J. Comput. Appl. Math.* **124** 281–302.

QI, Z., CUI, Y., LIU, Y. and PANG, J.-S. (2019). Estimation of individualized decision rules based on an optimized covariate-dependent equivalent of random outcomes. *SIAM J. Optim.* **29** 2337–2362. MR4008648 https://doi.org/10.1137/18M1190975

QI, Z., PANG, J.-S. and LIU, Y. (2023). On robustness of individualized decision rules. *J. Amer. Statist. Assoc.* **118** 2143–2157. MR4646632 https://doi.org/10.1080/01621459.2022.2038180

READ, J. and WILLIAMS, J. (2019). Positive and negative effects of antipsychotic medication: An international online survey of 832 recipients. *Curr. Drug. Saf.* **14** 173–181. https://doi.org/10.2174/1574886314666190301152734

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

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

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

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

SHAW, P. A. (2018). Use of composite outcomes to assess risk-benefit in clinical trials. *Clin. Trials* **15** 352–358. https://doi.org/10.1177/1740774518784010

SPIELMANS, G. I., SPENCE-SING, T. and PARRY, P. (2020). Duty to warn: Antidepressant black box suicidality warning is empirically justified. *Front Psychiatry* **11** 18. https://doi.org/10.3389/fpsyt.2020.00018

STROUP, T. S. and GRAY, N. (2018). Management of common adverse effects of antipsychotic medications. *World Psychiatry* **17** 341–356. https://doi.org/10.1002/wps.20567

STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., BYERLY, M. J., GLICK, I. D., CANIVE, J. M., MCGEE, M. F., SIMPSON, G. M., STEVENS, M. C. et al. (2003). The national institute of mental health clinical antipsychotic trials of intervention effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophr. Bull.* **29** 15–31. https://doi.org/10.1093/oxfordjournals.schbul.a006986

TAO, P. D. and AN, L. T. H. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Math. Vietnam.* **22** 289–355. MR1479751

THALL, P. F., SUNG, H.-G. and ESTEY, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Amer. Statist. Assoc.* **97** 29–39. MR1947271 https://doi.org/10.1198/016214502753479202

WANG, Y., FU, H. and ZENG, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: With an application to treating type 2 diabetes patients with insulin therapies. *J. Amer. Statist. Assoc.* **113** 1–13. MR3803435 https://doi.org/10.1080/01621459.2017.1303386

WATANABE, H., MARTINI, A. G., BROWN, E. A., LIANG, X., MEDRANO, S., GOTO, S., NARITA, I., AREND, L. J., SEQUEIRA-LOPEZ, M. L. S. et al. (2021). Inhibition of the renin-angiotensin system causes concentric hypertrophy of renal arterioles in mice and humans. *JCI Insight* **6** e154337.

WATKINS, C. J. C. H. (1989). Learning from delayed rewards (Ph.D. thesis).

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. MR3646564 https://doi.org/10.1080/01621459.2015.1093947

ZHU, S., SHEN, W., FU, H. and QU, A. (2024). Supplement to "Risk-aware restricted outcome learning for individualized treatment regimes of schizophrenia." https://doi.org/10.1214/23-AOAS1836SUPPA, https://doi.org/10.1214/23-AOAS1836SUPPB

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1337–1359
https://doi.org/10.1214/23-AOAS1837

PRIVACY-PRESERVING, COMMUNICATION-EFFICIENT, AND TARGET-FLEXIBLE HOSPITAL QUALITY MEASUREMENT

BY LARRY HAN[1,a], YIGE LI[2,b], BIJAN NIKNAM[3,c] AND JOSÉ R. ZUBIZARRETA[4,d]

[1]*Department of Health Sciences, Northeastern University and Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*lar.han@northeastern.edu*

[2]*Department of Biostatistics, Harvard T.H. Chan School of Public Health and Harvard CAUSALab,* [b]*yigeli@g.harvard.edu*

[3]*Department of Health Care Policy, Harvard Medical School,* [c]*bniknam1@jh.edu*

[4]*Departments of Health Care Policy, Biostatistics, Statistics, and Harvard CAUSALab,* [d]*zubizarreta@hcp.med.harvard.edu*

Accurate hospital performance measurement is important to both patients and providers but is challenging due to case-mix heterogeneity, differences in treatment guidelines, and data privacy regulations that preclude the sharing of individual patient data. Motivated to overcome these issues in the setting of hospital quality measurement, we develop a federated causal inference framework. We devise a doubly robust estimator of the mean potential outcome in a target population and show that it is consistent even when some models are misspecified. To enable real-world use, our proposed algorithm is privacy-preserving (requiring only summary statistics to be shared between hospitals) and communication-efficient (requiring only one round of communication between hospitals). We show that our estimator has good finite sample properties in simulation studies. We investigate the quality of hospital care provided by a diverse set of 51 candidate Cardiac Centers of Excellence, as measured by 30-day mortality and length of stay for acute myocardial infarction (AMI) patients. We find that our proposed federated global estimator improves the precision of treatment effect estimates by 34% to 86%, compared to using data from the target hospital alone. This precision gain results in qualitatively different conclusions about the estimated effect of percutaneous coronary intervention (PCI), compared to medical management (MM) in 43% (22 of 51) of hospitals.

REFERENCES

AUSTIN, P. C., ALTE, D. A. et al. (2003). Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: Should we be analyzing cardiovascular outcomes data differently? *Am. Heart J.* **145** 27–35.

BENJAMIN, E. J., BLAHA, M. J., CHIUVE, S. E. et al. (2017). Heart disease and stroke statistics-2017 update: A report from the American Heart Association. *Circulation* **135** e146–e603.

BRAUNWALD, E. (2014). The ten advances that have defined modern cardiology. *Trends Cardiovasc Med* **24** 179–183. https://doi.org/10.1016/j.tcm.2014.05.005

CMS (2020). Provider of Services Current Files.

CMS (2021). 2021 Condition-Specific Mortality Measures Updates and Specifications Report: Acute Myocardial Infarction- Version 15.0 Report, Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (YNHHSC/CORE) on behalf of Centers for Medicare & Medicaid Services (CMS).

DUAN, R., NING, Y., WANG, S., LINDSAY, B. G., CARROLL, R. J. and CHEN, Y. (2020). A fast score test for generalized mixture models. *Biometrics* **76** 811–820. MR4151850 https://doi.org/10.1111/biom.13204

GEORGE, E. I., ROČKOVÁ, V., ROSENBAUM, P. R., SATOPÄÄ, V. A. and SILBER, J. H. (2017). Mortality rate estimation and standardization for public reporting: Medicare's Hospital Compare. *J. Amer. Statist. Assoc.* **112** 933–947. MR3735351 https://doi.org/10.1080/01621459.2016.1276021

HAN, L., HOU, J., CHO, K., DUAN, R. and CAI, T. (2021). Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects. ArXiv preprint. Available at arXiv:2112.09313.

HAN, L., LI, Y., NIKNAM, B. and ZUBIZARRETA, J. R (2024). Supplement to "Privacy-preserving, communication-efficient, and target-flexible hospital quality measurement." https://doi.org/10.1214/23-AOAS1837SUPPA, https://doi.org/10.1214/23-AOAS1837SUPPB

Key words and phrases. Causal inference, data integration, federated learning, quality measurement.

HAN, P. and WANG, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika* **100** 417–430. MR3068443 https://doi.org/10.1093/biomet/ass087
KALBFLEISCH, J. D. and WOLFE, R. A. (2013). On monitoring outcomes of medical providers. *Stat. Biosci.* **5** 286–302.
KEELE, L. J., BEN-MICHAEL, E., FELLER, A., KELZ, R. and MIRATRIX, L. (2023). Hospital quality risk standardization via approximate balancing weights. *Ann. Appl. Stat.* **17** 901–928. MR4582697 https://doi.org/10.1214/22-aoas1629
KHATANA, S. A. M., NATHAN, A. S., DAYOUB, E. J., GIRI, J. and GROENEVELD, P. W. (2019). Centers of excellence designations, clinical outcomes, and characteristics of hospitals performing percutaneous coronary interventions. *JAMA Intern. Med.* **179** 1138–1140. https://doi.org/10.1001/jamainternmed.2019.0567
KRUMHOLZ, H. M., WANG, Y., MATTERA, J. A., WANG, Y., HAN, L. F., INGBER, M. J., ROMAN, S. and NORMAND, S. L. (2006). An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation* **113** 1683–92.
LASATER, K. B., MCHUGH, M. D., ROSENBAUM, P. R., AIKEN, L. H., SMITH, H. L., REITER, J. G., NIKNAM, B. A., HILL, A. S., HOCHMAN, L. L. et al. (2021). Evaluating the costs and outcomes of hospital nursing resources: A matched cohort study of patients with common medical conditions. *J. Gen. Intern. Med.* **36** 84–91. https://doi.org/10.1007/s11606-020-06151-z
LI, S., CAI, T. and DUAN, R. (2023). Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *Ann. Appl. Stat.* **17** 2970–2992. MR4661684 https://doi.org/10.1214/23-aoas1747
LI, T., SAHU, A. K., TALWALKAR, A. and SMITH, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **37** 50–60.
LONGFORD, N. T. (2020). Performance assessment as an application of causal inference. *J. Roy. Statist. Soc. Ser. A* **183** 1363–1385. MR4157817
MCDERMOTT, K. W., ELIXHAUSER, A. and SUN, R. (2017). Trends in Hospital Inpatient Stays in the United States, 2005–2014. *HCUP Statistical Brief* **225**.
NORMAND, S.-L. T., ASH, A. S., FIENBERG, S. E., STUKEL, T. A., UTTS, J. and LOUIS, T. A. (2016). League tables for hospital comparisons. *Annu. Rev. Stat. Appl.* **3** 21–50.
PATEL, M. R., CALHOON, J. H., DEHMER, G. J., GRANTHAM, J. A., MADDOX, T. M., MARON, D. J. and SMITH, P. K. (2017). ACC/AATS/AHA/ASE/ASNC/SCAI/SCCT/STS 2017 appropriate use criteria for coronary revascularization in patients with stable ischemic heart disease: A report of the American college of cardiology appropriate use criteria task force, American association for thoracic surgery, American heart association, American society of echocardiography, American society of nuclear cardiology, society for cardiovascular angiography and interventions, society of cardiovascular computed tomography, and society of thoracic surgeons. *J. Am. Coll. Cardiol.* **69** 2212–2241. https://doi.org/10.1016/j.jacc.2017.02.001
QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619
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. 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
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.
SHAW, F. E., ASOMUGHA, C. N., CONWAY, P. H. and REIN, A. S. (2014). The patient protection and affordable care act: Opportunities for prevention and public health. *Lancet* **384** 75–82.
SILBER, J. H., ARRIAGA, A. F., NIKNAM, B. A., HILL, A. S., ROSS, R. N. and ROMANO, P. S. (2018). Failure-to-rescue after acute myocardial infarction. *Med. Care* **56** 416–423. https://doi.org/10.1097/MLR.0000000000000904
SILBER, J. H., ROSENBAUM, P. R., BRACHET, T. J., ROSS, R. N., BRESSLER, L. J., EVEN-SHOSHAN, O., LORCH, S. A. and VOLPP, K. G. (2010). The hospital compare mortality model and the volume-outcome relationship. *Health Serv. Res.* **45** 1148–1167. https://doi.org/10.1111/j.1475-6773.2010.01130.x
SILBER, J. H., ROSENBAUM, P. R., NIKNAM, B. A., ROSS, R. N., REITER, J. G., HILL, A. S., HOCHMAN, L. L., BROWN, S. E., ARRIAGA, A. F. et al. (2020). Comparing outcomes and costs of surgical patients treated at major teaching and nonteaching hospitals: A national matched analysis. *Ann. Surg.* **271** 412–421.
SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., MUKHERJEE, N., SAYNISCH, P. A., EVEN-SHOSHAN, O. et al. (2014a). Template matching for auditing hospital cost and quality. *HSR* **49** 1446–74.
SILBER, J. H., ROSENBAUM, P. R., ROSS, R. N., LUDWIG, J. M., WANG, W., NIKNAM, B. A., SAYNISCH, P. A., EVEN-SHOSHAN, O., KELZ, R. R. et al. (2014b). A hospital-specific template for benchmarking its cost and quality. *Health Serv. Res.* **49** 1475–97.

SPLAWA-NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. D'abrowska and T. P. Speed. (1923) *Annals of Agricultural Sciences* 1–51. MR1092986

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1360–1377
https://doi.org/10.1214/23-AOAS1838

MASH: MEDIATION ANALYSIS OF SURVIVAL OUTCOME AND HIGH-DIMENSIONAL OMICS MEDIATORS WITH APPLICATION TO COMPLEX DISEASES

BY SUNYI CHI[1,a], CHRISTOPHER R. FLOWERS[2,e], ZIYI LI[1,b], XUELIN HUANG[1,c] AND PENG WEI[1,d]

[1]*Department of Biostatistics, University of Texas MD Anderson Cancer Center,* [a]*schi@mdanderson.org,* [b]*zli16@mdanderson.org,* [c]*xlhuang@mdanderson.org,* [d]*pwei2@mdanderson.org*

[2]*Department of Lymphoma, University of Texas MD Anderson Cancer Center,* [e]*crflowers@mdanderson.org*

Environmental exposures, such as cigarette smoking, influence health outcomes through intermediate molecular phenotypes, such as the methylome, transcriptome, and metabolome. Mediation analysis is a useful tool for investigating the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposures and health outcomes. However, little work has been done on mediation analysis when the mediators are high-dimensional and the outcome is a survival endpoint, and none of it has provided a robust measure of total mediation effect. To this end, we propose an estimation procedure for Mediation Analysis of Survival outcome and High-dimensional omics mediators (MASH), based on a second-moment-based measure of total mediation effect for survival data analogous to the R^2 measure in a linear model. In addition, we propose a three-step mediator selection procedure to mitigate potential bias induced by nonmediators. Extensive simulations showed good performance of MASH in estimating the total mediation effect and identifying true mediators. By applying MASH to the metabolomics data of 1919 subjects in the Framingham Heart Study, we identified five metabolites as mediators of the effect of cigarette smoking on coronary heart disease risk (total mediation effect, 51.1%) and two metabolites as mediators between smoking and risk of cancer (total mediation effect, 50.7%). Application of MASH to a diffuse large B-cell lymphoma genomics data set identified copy-number variations for eight genes as mediators between the baseline International Prognostic Index score and overall survival.

REFERENCES

ALTMAN, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Amer. Statist.* **46** 175–185. MR1183070 https://doi.org/10.2307/2685209

AMERICAN CANCER SOCIETY (2021). *Cancer Preventio & Early Detection Facts & Figures 2021–2022*. American Cancer Society, Atlanta, Ga.

BALDARI, C. T. (2016). S1PR2 deficiency in DLBCL: A FOXy connection. *Blood* **127** 1380–1381.

BARRANS, S. L., FENTON, J. A. L., BANHAM, A., OWEN, R. G. and JACK, A. S. (2004). Strong expression of FOXP1 identifies a distinct subset of diffuse large B-cell lymphoma (DLBCL) patients with poor outcome. *Blood* **104** 2933–2935. https://doi.org/10.1182/blood-2004-03-1209

BENOWITZ, N. L. (1996). Cotinine as a biomarker of environmental tobacco smoke exposure. *Epidemiol. Rev.* **18** 188–204. https://doi.org/10.1093/oxfordjournals.epirev.a017925

CAVUS, E., KARAKAS, M., OJEDA, F. M., KONTTO, J., VERONESI, G., FERRARIO, M. M. et al. (2019). Association of circulating metabolites with risk of coronary heart disease in a European population: Results from the biomarkers for cardiovascular risk assessment in Europe (BiomarCaRE) Consortium. *JAMA Cardiol.* **4** 1270–1279.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2020). *Health Effects of Cigarette Smoking*.

CHI, S., FLOWERS, C. R., LI, Z., HUANG, X. and WEI, P. (2024). Supplement to "MASH: Mediation analysis of survival outcome and high-dimensional omics mediators with application to complex diseases." https://doi.org/10.1214/23-AOAS1838SUPP

Key words and phrases. High-dimensional mediators, mediation analysis, survival analysis, total mediation effect, variable selection.

CLARKE, M. B., BEZABEH, D. Z. and HOWARD, C. T. (2006). Determination of carbohydrates in tobacco products by liquid chromatography-mass spectrometry/mass spectrometry: A comparison with ion chromatography and application to product discrimination. *J. Agric. Food Chem.* **54** 1975–1981.

CROSS, A. J., BOCA, S., FREEDMAN, N. D., CAPORASO, N. E., HUANG, W. Y., SINHA, R., SAMPSON, J. N. and MOORE, S. C. (2014). Metabolites of tobacco smoking and colorectal cancer risk. *Carcinogenesis* **35** 1516–1522.

DAI, J., WANG, H., DONG, Y., ZHANG, Y. and WANG, J. (2013). Bile acids affect the growth of human cholangiocarcinoma via NF-kB pathway. *Cancer Investig.* **31** 111–120.

DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N. and VANSTEELANDT, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71** 1–14. MR3335344 https://doi.org/10.1111/biom.12248

DINICOLANTONIO, J. J., LAVIE, C. J., FARES, H., MENEZES, A. R. and O'KEEFE, J. H. (2013). L-carnitine in the secondary prevention of cardiovascular disease: Systematic review and meta-analysis. *Mayo Clin. Proc.* **88** 544–551.

ELLARD, G. A., DE WAARD, F. and KEMMEREN, J. M. (1995). Urinary nicotine metabolite excretion and lung cancer risk in a female cohort. *Br. J. Cancer* **72** 788–791.

FAIRCHILD, A. J., MACKINNON, D. P., TABORGA, M. P. and TAYLOR, A. B. (2009). R2 effect-size measures for mediation analysis. *Behav. Res. Methods* **41** 486–498.

FAN, J., FENG, Y. and WU, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect.* **6** 70–86. IMS, Beachwood, OH. MR2798512

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

GOSSETT, L. K., JOHNSON, H. M., PIPER, M. E., FIORE, M. C., BAKER, T. B. and STEIN, J. H. (2009). Smoking intensity and lipoprotein abnormalities in active smokers. *J. Clin. Lipidol.* **3** 372–378. https://doi.org/10.1016/j.jacl.2009.10.008

HOWARD, B. V. and WYLIE-ROSETT, J. (2002). Sugar and cardiovascular disease: A statement for healthcare professionals from the committee on nutrition of the council on nutrition, physical activity, and metabolism of the American heart association. *Circulation* **106** 523–527.

HUANG, Y. T. and YANG, H. I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* **28** 370–8.

INTERNATIONAL NON-HODGKIN'S LYMPHOMA PROGNOSTIC FACTORS PROJECT (INHLPFP) (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *N. Engl. J. Med.* **329** 987–994.

KENT, J. T. and O'QUIGLEY, J. (1988). Measures of dependence for censored survival data. *Biometrika* **75** 525–534. MR0967592 https://doi.org/10.1093/biomet/75.3.525

KÜHN, T., STEPIEN, M., LÓPEZ-NOGUEROLES, M., DAMMS-MACHADO, A., SOOKTHAI, D., JOHNSON, T., ROCA, M., HÜSING, A., MALDONADO, S. G. et al. (2020). Prediagnostic plasma bile acid levels and colon cancer risk: A prospective study. *J. Natl. Cancer Inst.* **112** 516–524.

KUNUTSOR, S. K., SPEE, J. M., KIENEKER, L. M., GANSEVOORT, R. T., DULLAART, R. P. F., VOERMAN, A. J., TOUW, D. J. and BAKKER, S. J. L. (2018). Self-reported smoking, urine cotinine, and risk of cardiovascular disease: Findings from the PREVEND (prevention of renal and vascular end-stage disease) prospective cohort study. *J. Amer. Heart Assoc.* **7** e008726.

LI, Y., ZHANG, D., HE, Y. et al. (2017). Investigation of novel metabolites potentially involved in the pathogenesis of coronary heart disease using a UHPLC-QTOF/MS-based metabolomics approach. *Sci. Rep.* **7** 15357.

LINDENBERGER, U. and PÖTTER, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychol. Methods* **3** 218–230.

LIU, Z., SHEN, J., BARFIELD, R., SCHWARTZ, J., BACCARELLI, A. A. and LIN, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *J. Amer. Statist. Assoc.* **117** 67–81. MR4399068 https://doi.org/10.1080/01621459.2021.1914634

LUO, C., FA, B., YAN, Y., WANG, Y., ZHOU, Y., ZHANG, Y. and YU, Z. (2020). High-dimensional mediation analysis in survival models. *PLoS Comput. Biol.* **16** e1007768.

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

MAZZILLI, K. M., MCCLAIN, K. M., LIPWORTH, L., PLAYDON, M. C., SAMPSON, J. N., CLISH, C. B., GERSZTEN, R. E., FREEDMAN, N. D. and MOORE, S. C. (2020). Identification of 102 correlations between serum metabolites and habitual diet in a metabolomics study of the prostate, lung, colorectal, and ovarian cancer trial. *J. Nutr.* **150** 694–703.

MUNDRA, P. A., BARLOW, C. K., NESTEL, P. J., BARNES, E. H., KIRBY, A., THOMPSON, P., SULLIVAN, D. R., ALSHEHRY, Z. H., MELLETT, N. A. et al. (2018). Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention. *JCI Insight* **3** e121326.

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

REDDY, A., ZHANG, J., DAVIS, N. S., MOFFITT, A. B., LOVE, C. L., WALDROP, A., LEPPA, S., PASANEN, A., MERIRANTA, L. et al. (2017). Genetic and functional drivers of diffuse large B cell lymphoma. *Cell* **171** 481–494.e15.

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

ROCHE, H. M. and GIBNEY, M. J. (2000). Effect of long-chain n-3 polyunsaturated fatty acids on fasting and postprandial triacylglycerol metabolism. *Am. J. Clin. Nutr.* **71** 232S–7S.

ROYSTON, P. (2006). Explained variation for survival models. *Stata J.* **6** 83–96.

SAMPSON, J. N., BOCA, S. M., MOORE, S. C. and HELLER, R. (2018). FWER and FDR control when testing multiple mediators. *Bioinformatics* **34** 2418–24.

SCHEMPER, M. and STARE, J. (1996). Explained variation in survival analysis. *Stat. Med.* **15** 1999–2012.

SCHMITZ, R., WRIGHT, G. W., HUANG, D. W., STAUDT, L. et al. (2018). Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.* **378** 1396–1407.

SHI, B., HUANG, X. and WEI, P. (2022). Comparison of effect size measures for mediation analysis of survival outcomes with application to the framingham heart study. Available at arXiv:2205.03303.

SPLANSKY, G. L., COREY, D., YANG, Q., ATWOOD, L. D., CUPPLES, L. A., BENJAMIN, E. J., D'AGOSTINO SR, R. B., FOX, C. S., LARSON, M. G. et al. (2007). The third generation cohort of the national heart, lung, and blood institute's framingham heart study: Design, recruitment, and initial examination. *Amer. J. Epidemiol.* **165** 1328–1335.

SUBBAIAH, P. V., JIANG, X. C., BELIKOVA, N. A., AIZEZI, B., HUANG, Z. H. and REARDON, C. A. (2012). Regulation of plasma cholesterol esterification by sphingomyelin: Effect of physiological variations of plasma sphingomyelin on lecithin-cholesterol acyltransferase activity. *Biochim. Biophys. Acta* **1821** 908–913.

TEIN, J.-Y. and MACKINNON, D. P. (2003). Estimating mediated effects with survival data. In *New Developments on Psychometrics* (H. Yanai, A. O. Rikkyo, K. Shigemasu, Y. Kano and J. J. Meulman, eds.) 405–412. Springer, Tokyo, Japan.

THONG, A. E., PETRUZELLA, S., ORLOW, I., ZABOR, E. C., EHDAIE, B., OSTROFF, J. S., BOCHNER, B. H. and BARNES, H. F. (2016). Accuracy of self-reported smoking exposure among bladder cancer patients undergoing surveillance at a tertiary referral center. *Eur. Urol. Focus* **2** 441–444.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2010). *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis For Smoking-Attributable Disease: A Report of the Surgeon General*. U.S. Dept. Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2020). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General*. U.S. Dept. Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta. 2014 [accessed 2020 January 27].

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2020). *A Report of the Surgeon General. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease*. U.S. Dept. Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta. 2010 [accessed 2020 January 27].

VANDERWEELE, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology* **22** 582–585. https://doi.org/10.1097/EDE.0b013e31821db37e

VANDERWEELE, T. J. (2016). Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* **37** 17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402

VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiol. Methods* **2** 95–115.

WANG, H. H., GARRUTI, G., LIU, M., PORTINCASA, P. and WANG, D. Q. (2017). Cholesterol and lipoprotein metabolism and atherosclerosis: Recent advances in reverse cholesterol transport. *Ann. Hepatol.* **16** s27–s42.

WANG, Z., ZHU, C., NAMBI, V., MORRISON, A. C., FOLSOM, A. R., BALLANTYNE, C. M., BOERWINKLE, E. and YU, B. (2019). Metabolomic pattern predicts incident coronary heart disease. *Arterioscler. Thromb. Vasc. Biol.* **39** 1475–1482.

WHINCUP, P. H., GILG, J. A., EMBERSON, J. R., JARVIS, M. J., FEYERABEND, C., BRYANT, A., WALKER, M. and COOK, D. G. (2004). Passive smoking and risk of coronary heart disease and stroke: Prospective study with cotinine measurement. *BMJ, Br. Med. J. (Clin. Res. Ed.)* **329** 200–205.

XU, P. P., HUO, Y. J. and ZHAO, W. L. (2022). All roads lead to targeted diffuse large B-cell lymphoma approaches. *Cancer Cell* **40** 131–133.

XU, T., HOLZAPFEL, C., DONG, X. et al. (2013). Effects of smoking and smoking cessation on human serum metabolite profile: Results from the KORA cohort study. *BMC Med.* **11** 60.

YANG, T., NIU, J., CHEN, H. and WEI, P. (2021). Estimation of mediation effect for high-dimensional omics mediators. *BMC Bioinform.* **22** 414.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHAO, G., ZHANG, H., WANG, Y., GAO, X., LIU, H. and LIU, W. (2020). Effects of levocarnitine on cardiac function, urinary albumin, hs-CRP, BNP, and troponin in patients with coronary heart disease and heart failure. *Ell. Kardiol. Epitheōr.*: *HJC = Hellenike kardiologike epitheorese* **61** 99–102.

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

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1378–1403
https://doi.org/10.1214/23-AOAS1839

FLEXIBLE MULTIVARIATE SPATIOTEMPORAL HAWKES PROCESS MODELS OF TERRORISM

BY MIKYOUNG JUN[1,a] AND SCOTT COOK[2,b]

[1]*Department of Mathematics, College of Natural Sciences and Mathematics, University of Houston,* [a]*mjun@central.uh.edu*
[2]*Department of Political Science, Bush School of Government and Public Service, Texas A&M University,* [b]*sjcook@tamu.edu*

We develop flexible multivariate spatiotemporal Hawkes process models to analyze patterns of terrorism. Previous applications of point process methods to political violence data mainly utilize temporal Hawkes process models, neglecting spatial variation in these attack patterns. This limits what can be learned from these models, as any effective counter-terrorism strategy requires knowledge on both when and where attacks are likely to occur. Even the existing work on spatiotemporal Hawkes processes imposes restrictions on the triggering function that are not well-suited for terrorism data. Therefore, we generalize the structure of the spatiotemporal triggering function considerably, allowing for nonseparability, nonstationarity, and cross-triggering (across multiple terror groups). To demonstrate the utility of our models, we analyze two samples of real-world terrorism data: Afghanistan (2002–2013) as a univariate analysis and Nigeria (2009–2017) as a bivariate analysis. Jointly, these two studies demonstrate that our generalized models outperform standard Hawkes process models, besting widely-used alternatives in overall model fit and revealing spatiotemporal patterns that are, by construction, masked in these models (e.g., increasing dispersion in cross-triggering over time).

REFERENCES

BOVE, V. and BÖHMELT, T. (2016). Does immigration induce terrorism? *J. Polit.* **78** 572–588.

BRAITHWAITE, A. and LI, Q. (2007). Transnational terrorism hot spots: Identification and impact evaluation. *Confl. Manage. Peace Sci.* **24** 281–296.

BRÉMAUD, P. and MASSOULIÉ, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Probab.* **24** 1563–1588. MR1411506 https://doi.org/10.1214/aop/1065725193

CHAKRABORTY, A. and GELFAND, A. E. (2010). Analyzing spatial point patterns subject to measurement error. *Bayesian Anal.* **5** 97–122. MR2596437 https://doi.org/10.1214/10-BA504

CHEN, F. and HALL, P. (2013). Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling. *J. Appl. Probab.* **50** 1006–1024. MR3161370 https://doi.org/10.1239/jap/1389370096

CHEN, F. and HALL, P. (2016). Nonparametric estimation for self-exciting point processes—a parsimonious approach. *J. Comput. Graph. Statist.* **25** 209–224. MR3474044 https://doi.org/10.1080/10618600.2014.1001491

CHEN, J., HAWKES, A. G., SCALAS, E. and TRINH, M. (2018). Performance of information criteria for selection of Hawkes process models of financial data. *Quant. Finance* **18** 225–235. MR3750732 https://doi.org/10.1080/14697688.2017.1403140

CHEN, Y. (2016). Multivariate Hawkes processes and their simulations. Available at https://www.math.fsu.edu/~ychen/research/multiHawkes.pdf.

CHENG, Y., DUNDAR, M. and MOHLER, G. (2018). A coupled ETAS-I^{2}GMM point process with applications to seismic fault detection. *Ann. Appl. Stat.* **12** 1853–1870. MR3852700 https://doi.org/10.1214/18-AOAS1134

COOK, S. J. and WEIDMANN, N. B. (2022). Race to the bottom: Spatial aggregation and event data. *Int. Interact.* **48** 471–491.

CRENSHAW, M. (2000). The psychology of terrorism: An agenda for the 21st century. *Political Psycholog.* **21** 405–420.

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

Key words and phrases. GTD, Hawkes processes, multivariate point process, spatiotemporal point patterns, terrorism.

DIGGLE, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed. *Monographs on Statistics and Applied Probability* **128**. CRC Press, Boca Raton, FL. MR3113855

DIGGLE, P. J., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statist. Sci.* **28** 542–563. MR3161587 https://doi.org/10.1214/13-STS441

ENDERS, W. and SANDLER, T. (2006). Distribution of transnational terrorism among countries by income class and geography after 9/11. *Int. Stud. Q.* **50** 367–393.

FANSHAWE, T. R. and DIGGLE, P. J. (2011). Spatial prediction in the presence of positional error. *Environmetrics* **22** 109–122. MR2843341 https://doi.org/10.1002/env.1062

FINDLEY, M. G. and YOUNG, J. K. (2012). Terrorism and civil war: A spatial and temporal approach to a conceptual problem. *Perspective Polit.* **10** 285–305.

INSTITUTE FOR ECONOMICS & PEACE (2020). Global Terrorism Index 2020: Measuring the impact of terrorism. Available at http://visionofhumanity.org/reports.

START (NATIONAL CONSORTIUM FOR THE STUDY OF TERRORISM AND RESPONSES TO TERRORISM) (2022). Global Terrorism Database 1970-2020 [data file] Available at https://www.start.umd.edu/gtd.

FUHRIMAN, C., MEDINA, R. M. and BREWER, S. (2017). A point process analysis of terror attacks in Afghanistan, 2002-2013. *Middle States Geogr.* **50** 50–63.

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475 https://doi.org/10.1198/016214502760047113

GONZÁLEZ, J. A., RODRÍGUEZ-CORTÉS, F. J., CRONIE, O. and MATEU, J. (2016). Spatio-temporal point process statistics: A review. *Spat. Stat.* **18** 505–544. MR3575505 https://doi.org/10.1016/j.spasta.2016.10.002

HEATON, M. J., BERRETT, C., PUGH, S., EVANS, A. and SLOAN, C. (2020). Modeling bronchiolitis incidence proportions in the presence of spatio-temporal uncertainty. *J. Amer. Statist. Assoc.* **115** 66–78. MR4078445 https://doi.org/10.1080/01621459.2019.1609480

HOLLISTER, J., SHAH, T., ROBITAILLE, A. L., BECK, M. W. and JOHNSON, M. (2021). elevatr: Access elevation data from various APIs. R package version 0.4.2.

ILHAN, F. and KOZAT, S. S. (2020). Modeling of spatio-temporal Hawkes processes with randomized kernels. *IEEE Trans. Signal Process.* **68** 4946–4958. MR4154070 https://doi.org/10.1109/TSP.2020.3019329

JANG, H. J., LEE, K. and LEE, K. (2019). Systemic risk in market microstructure of crude oil and gasoline futures prices: A Hawkes flocking model approach. *J. Futures Mark.* **40** 247–275.

JOHNSON, N., HITCHMAN, A., PHAN, D. and SMITH, L. (2018). Self-exciting point process models for political conflict forecasting. *European J. Appl. Math.* **29** 685–707. MR3819992 https://doi.org/10.1017/S095679251700033X

JUN, M. (2011). Non-stationary cross-covariance models for multivariate processes on a globe. *Scand. J. Stat.* **38** 726–747. MR2859747 https://doi.org/10.1111/j.1467-9469.2011.00751.x

JUN, M. and COOK, S. (2024). Supplement to "Flexible multivariate spatiotemporal Hawkes process models of terrorism." https://doi.org/10.1214/23-AOAS1839SUPPA, https://doi.org/10.1214/23-AOAS1839SUPPB

JUN, M., SCHUMACHER, C. and SARAVANAN, R. (2019). Global multivariate point pattern models for rain type occurrence. *Spat. Stat.* **31** 100355. MR3946994 https://doi.org/10.1016/j.spasta.2019.04.003

JUN, M. and STEIN, M. L. (2007). An approach to producing space-time covariance functions on spheres. *Technometrics* **49** 468–479. MR2394558 https://doi.org/10.1198/004017007000000155

LE, T. M. (2018). A multivariate Hawkes process with gaps in observations. *IEEE Trans. Inf. Theory* **64** 1800–1811. MR3766315 https://doi.org/10.1109/TIT.2017.2735963

LIU, X., CARTER, J., RAY, B. and MOHLER, G. (2021). Point process modeling of drug overdoses with heterogeneous and missing data. *Ann. Appl. Stat.* **15** 88–101. MR4255268 https://doi.org/10.1214/20-aoas1384

MARINEAU, J., PASCOE, H., BRAITHWAITE, A., FINDLEY, M. and YOUNG, J. (2020). The local geography of transnational terrorism. *Confl. Manage. Peace Sci.* **37** 350–381.

MOHLER, G., MCGRATH, E., BUNTAIN, C. and LAFREE, G. (2020). Hawkes binomial topic model with applications to coupled conflict-Twitter data. *Ann. Appl. Stat.* **14** 1984–2002. MR4194257 https://doi.org/10.1214/20-AOAS1352

MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

NEMETH, S. C., MAUSLEIN, J. A. and STAPLEY, C. (2014). The primacy of the local: Identifying terrorist hot spots using geographic information systems. *J. Polit.* **76** 304–317.

NEUMAYER, E. and PLÜMPER, T. (2010). Galton's problem and contagion in international terrorism along civilizational lines. *Confl. Manage. Peace Sci.* **27** 308–325.

NEWMAN, L. S. (2013). Do terrorist attacks increase closer to elections? *Terrorism Polit. Violence* **25** 8–28.

POLO, S. M. (2020). How terrorism spreads: Emulation and the diffusion of ethnic and ethnoreligious terrorism. *J. Confl. Resolut.* **64** 1916–1942.

PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531 https://doi.org/10.1214/11-AOAS513

PYTHON, A., BRANDSCH, J., ILLIAN, J. B., JONES-TODD, C. M. and BLANGIARDO, M. (2019a). Statistics and terrorism: Insights into lethality of terrorism through Bayesian modeling. *Wiley StatsRef*. https://doi.org/10.1002/9781118445112.stat08250

PYTHON, A., ILLIAN, J. B., JONES-TODD, C. M. and BLANGIARDO, M. (2019b). A Bayesian approach to modelling subnational spatial dynamics of worldwide non-state terrorism, 2010–2016. *J. Roy. Statist. Soc. Ser. A* **182** 323–344. MR3902646 https://doi.org/10.1111/rssa.12384

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

REINHART, A. and GREENHOUSE, J. (2018). Self-exciting point processes with spatial covariates: Modelling the dynamics of crime. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1305–1329. MR3873709 https://doi.org/10.1111/rssc.12277

RENNER, I. W., ELITH, J., BADDELEY, A., FITHIAN, W., HASTIE, T., PHILLIPS, S. J., POPOVIC, G. and WARTON, D. I. (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.* **6** 366–379.

ROUEFF, F. and VON SACHS, R. (2019). Time-frequency analysis of locally stationary Hawkes processes. *Bernoulli* **25** 1355–1385. MR3920375 https://doi.org/10.3150/18-bej1023

RUBIN, G. J., BREWIN, C. R., GREENBERG, N., HUGHES, J. H., SIMPSON, J. and WESSELY, S. (2007). Enduring consequences of terrorism: 7-month follow-up survey of reactions to the bombings in London on 7 July 2005. *Br. J. Psychiatry* **190** 350–356.

SANDLER, T. (2014). The analytical study of terrorism: Taking stock. *J. Peace Res.* **51** 257–271.

SANDLER, T. and ENDERS, W. (2008). Economic consequences of terrorism in developed and developing countries. *Terror. Econ. Dev. Polit. Openness* **17**.

SCHNEIDER, F., BRÜCK, T. and MEIERRIEKS, D. (2015). The economics of counterterrorism: A survey. *J. Econ. Surv.* **29** 131–157.

SCHOENBERG, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc.* **98** 789–795. MR2055487 https://doi.org/10.1198/016214503000000710

SCHOENBERG, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statist. Sinica* **26** 861–879. MR3497774

SCHOENBERG, F. P., BRILLINGER, D. R. and GUTTORP, P. M. (2002). Encyclopedia of Environmetrics **3** 1573–1577 Point processes, spatial-temporal. Wiley, New York.

SCHOENBERG, F. P., HOFFMANN, M. and HARRIGAN, R. J. (2019). A recursive point process model for infectious diseases. *Ann. Inst. Statist. Math.* **71** 1271–1287. MR3993533 https://doi.org/10.1007/s10463-018-0690-9

SIEBENECK, L. K., MEDINA, R. M., YAMADA, I. and HEPNER, G. F. (2009). Spatial and temporal analyses of terrorist incidents in Iraq, 2004–2006. *Stud. Confl. Terrorism* **32** 591–610.

SOLIMAN, H., ZHAO, L., HUANG, Z., PAUL, S. and XU, K. S. (2022). The multivariate community Hawkes model for dependent relational events in continuous-time networks. In *Proceedings of the* 39*th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, eds.). *Proceedings of Machine Learning Research* **162** 20329–20346. PMLR.

STEIN, M. and HIRSHBERG, A. (1999). Medical consequences of terrorism. The conventional weapon threat. *Surg. Clin. North Amer.* **79** 1537–1552. https://doi.org/10.1016/s0039-6109(05)70091-8

STEIN, M. L. (2005a). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. MR2156840 https://doi.org/10.1198/016214504000000854

STEIN, M. L. (2005b). Nonstationary spatial covariance functions Technical Report No. 21 Center for Integrating Statistical and Environmental Science, The Univ. Chicago.

TENCH, S., FRY, H. and GILL, P. (2016). Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army. *European J. Appl. Math.* **27** 377–402. MR3491504 https://doi.org/10.1017/S0956792515000686

TURK, A. T. (2004). Sociology of terrorism. *Annu. Rev. Sociol.* **30** 271–286.

WANG, S. (2021). Self-exciting point process for modeling terror attack data Ph.D. thesis Wilfrid Laurier Univ.

WEIDMANN, N. B. (2015). On the accuracy of media-based conflict event data. *J. Confl. Resolut.* **59** 1129–1149.

WHITE, G., PORTER, M. D. and MAZEROLLE, L. (2013). Terrorism risk, resilience and volatility: A comparison of terrorism patterns in three southeast Asian countries. *J. Quant. Criminol.* **29** 295–320.

YUAN, B., LI, H., BERTOZZI, A. L., BRANTINGHAM, P. J. and PORTER, M. A. (2019). Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM J. Math. Data Sci.* **1** 356–382. MR3975150 https://doi.org/10.1137/18M1226993

ZHU, L., COOK, S. J. and JUN, M. (2021). The promise and perils of point process models of political events. Available at arXiv:2108.12566v1.

ZHU, L., YANG, J., JUN, M. and COOK, S. (2022). On minimum contrast method for multivariate spatial point processes. Available at arXiv:2208.07044 [stat.ME].

ZIMMERMAN, D. L., FANG, X., MAZUMDAR, S. and RUSHTON, G. (2007). Modeling the probability distribution of positional errors incurred by residential address geocoding. *Int. J. Health Geogr.* **6** 1–16.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1404–1420
https://doi.org/10.1214/23-AOAS1840

HIERARCHICAL DEPENDENCE MODELING FOR THE ANALYSIS OF LARGE INSURANCE CLAIMS DATA

BY TING FUNG MA[1,a], YIZHOU CAI[1,b], PENG SHI[2,c] AND JUN ZHU[3,d]

[1]*Department of Statistics, University of South Carolina,* [a]*tingfung@mailbox.sc.edu,* [b]*yizhouc@email.sc.edu*
[2]*Department of Risk and Insurance, Wisconsin School of Business, University of Wisconsin-Madison,* [c]*pshi@bus.wisc.edu*
[3]*Department of Statistics, University of Wisconsin-Madison,* [d]*jzhu@stat.wisc.edu*

Extreme weather events associated with climate change have caused significant damages. In particular, hail storms damage millions of properties in the U.S. and result in billion-dollar insured losses each year in the recent decade. To facilitate the insurance claims management operations in insurance companies, we construct a hierarchical dependence model, which accommodates the complex dependence within and between the outcomes of interests including the propensity of filing a claim, time to report a claim, and the claim amount. The storm-specific and property-specific characteristics are incorporated through marginal models, such as generalized linear models and survival analysis models. The dependence within the hail event is captured by spatial factor copula, while the dependence between different outcomes is captured by bivariate copula. For parameter estimation we develop a two-step procedure that first maximizes the marginal likelihood function and then maximizes the pairwise likelihood, which ensures computational feasibility for big data. We apply this modeling framework to analyze a large dataset involving hail storms in Colorado from 2011 to 2015 impacting hundreds of thousands of insured properties and demonstrate that the predictive performance can be improved by our proposed methodology.

REFERENCES

ALLEN, J. T., GIAMMANCO, I. M., KUMJIAN, M. R., JURGEN PUNGE, H., ZHANG, Q., GROENEMEIJER, P., KUNZ, M. and ORTEGA, K. (2020). Understanding hail in the Earth system. *Rev. Geophys.* **58** e2019RG000665. https://doi.org/10.1029/2019RG000665

BAI, Y., KANG, J. and SONG, P. X.-K. (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics* **70** 661–670. MR3261785 https://doi.org/10.1111/biom.12199

BAI, Y., SONG, P. X.-K. and RAGHUNATHAN, T. E. (2012). Joint composite estimating functions in spatiotemporal models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 799–824. MR2988907 https://doi.org/10.1111/j.1467-9868.2012.01035.x

BEVILACQUA, M., GAETAN, C., MATEU, J. and PORCU, E. (2012). Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *J. Amer. Statist. Assoc.* **107** 268–280. MR2949358 https://doi.org/10.1080/01621459.2011.646928

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

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. Revised reprint of the 1991 edition, A Wiley-Interscience Publication. MR1239641 https://doi.org/10.1002/9781119115151

GAO, L. and SHI, P. (2022). Leveraging high-resolution weather information to predict hail damage claims: A spatial point process for replicated point patterns. *Insurance Math. Econom.* **107** 161–179. MR4477473 https://doi.org/10.1016/j.insmatheco.2022.08.006

GAO, X. and SONG, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc.* **105** 1531–1540. MR2796569 https://doi.org/10.1198/jasa.2010.tm09414

Key words and phrases. Composite likelihood, copula, non-Gaussian data, nonstationary process, replicated data, two-step estimation.

GENZ, A. and BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics* **195**. Springer, Dordrecht. MR2840595 https://doi.org/10.1007/978-3-642-01689-9

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

HANSEN, B. E. and LEE, S. (2019). Asymptotic theory for clustered samples. *J. Econometrics* **210** 268–290. MR3958406 https://doi.org/10.1016/j.jeconom.2019.02.001

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.* **94** 401–419. MR2167922 https://doi.org/10.1016/j.jmva.2004.06.003

KO, V. and HJORT, N. L. (2019). Model robust inference with two-stage maximum likelihood estimation for copulas. *J. Multivariate Anal.* **171** 362–381. MR3907859 https://doi.org/10.1016/j.jmva.2019.01.004

KRUPSKII, P., HUSER, R. and GENTON, M. G. (2018). Factor copula models for replicated spatial data. *J. Amer. Statist. Assoc.* **113** 467–479. MR3803479 https://doi.org/10.1080/01621459.2016.1261712

LINDSAY, B. G., YI, G. Y. and SUN, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statist. Sinica* **21** 71–105. MR2796854

MA, T. F., CAI, Y., SHI, P. and ZHU, J. (2024). Supplement to "Hierarchical dependence modeling for the analysis of large insurance claims data." https://doi.org/10.1214/23-AOAS1840SUPPA, https://doi.org/10.1214/23-AOAS1840SUPPB

MA, T. F. and YAU, C. Y. (2016). A pairwise likelihood-based approach for changepoint detection in multivariate time series models. *Biometrika* **103** 409–421. MR3509895 https://doi.org/10.1093/biomet/asw002

NG, C. T. and JOE, H. (2014). Model comparison with composite likelihood information criteria. *Bernoulli* **20** 1738–1764. MR3263088 https://doi.org/10.3150/13-BEJ539

RAUPACH, T. H., MARTIUS, O., ALLEN, J. T., KUNZ, M., LASHER-TRAPP, S., MOHR, S., RASMUSSEN, K. L., TRAPP, R. J. and ZHANG, Q. (2021). The effects of climate change on hailstorms. *Nat. Rev. Earth Environ.* **2** 213–226. https://doi.org/10.1038/s43017-020-00133-9

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

SHI, P., FUNG, G. M. and DICKINSON, D. (2022). Assessing hail risk for property insurers with a dependent marked point process. *J. Roy. Statist. Soc. Ser. A* **185** 302–328. MR4384307 https://doi.org/10.1111/rssa.12754

SHI, P. and SHI, K. (2017). Territorial risk classification using spatially dependent frequency-severity models. *Astin Bull.* **47** 437–465. MR3654418 https://doi.org/10.1017/asb.2017.7

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

VARIN, C. and VIDONI, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92** 519–528. MR2202643 https://doi.org/10.1093/biomet/92.3.519

VERISK (2021). Understanding evolving hail risk Verisk Hail Report 1–13.

WANG, P., MA, T. F., BANDYOPADHYAY, D., TANG, Y. and ZHU, J. (2021). Composite likelihood inference for ordinal periodontal data with replicated spatial patterns. *Stat. Med.* **40** 5871–5893. MR4330584 https://doi.org/10.1002/sim.9160

ZHAO, Y. and JOE, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33** 335–356. MR2193979 https://doi.org/10.1002/cjs.5540330303

ZHAO, Z., SHI, P. and FENG, X. (2021). Knowledge learning of insurance risks using dependence models. *INFORMS J. Comput.* **33** 1177–1196. MR4311275 https://doi.org/10.1287/ijoc.2020.1005

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1421–1444
https://doi.org/10.1214/23-AOAS1841

FORECASTING U.S. INFLATION USING BAYESIAN NONPARAMETRIC MODELS

BY TODD E. CLARK[1,a], FLORIAN HUBER[2,b], GARY KOOP[3,c] AND MASSIMILIANO MARCELLINO[4,d]

[1]*Research Department, Federal Reserve Bank of Cleveland,* [a]*todd.clark@researchfed.org*
[2]*Department of Economics, University of Salzburg,* [b]*florian.huber@plus.ac.at*
[3]*Department of Economics, University of Strathclyde,* [c]*gary.koop@strath.ac.uk*
[4]*Department of Economics, Bocconi University, IGIER, and CEPR,* [d]*massimiliano.marcellino@unibocconi.it*

The relationship between inflation and predictors, such as unemployment, is potentially nonlinear with a strength that varies over time, and prediction errors may be subject to large, asymmetric shocks. Inspired by these concerns, we develop a model for inflation forecasting that is nonparametric both in the conditional mean and in the error using Gaussian and Dirichlet processes, respectively. We discuss how both these features may be important in producing accurate forecasts of inflation. In a forecasting exercise involving CPI inflation, we find that our approach has substantial benefits, both overall and in the left tail, with nonparametric modeling of the conditional mean being of particular importance.

REFERENCES

BABB, N. R. and DETMEISTER, A. K. (2017). Nonlinearities in the Phillips curve for the United States: Evidence using metropolitan data. Working Paper No. 2017-070. Board of Governors of the Federal Reserve System. https://doi.org/10.17016/FEDS.2017.070

BRAUN, R. (2021). The importance of supply and demand for oil prices: Evidence from non-Gaussianity. *Quant. Econ.* forthcoming.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32. https://doi.org/10.1023/A:1010933404324

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948. https://doi.org/10.2307/2669832

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

CLARK, T. E. (2011). Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility. *J. Bus. Econom. Statist.* **29** 327–341. MR2848507 https://doi.org/10.1198/jbes.2010.09248

CLARK, T. E., HUBER, F., KOOP, G. and MARCELLINO, M. (2024). Supplement to "Forecasting U.S. Inflation Using Bayesian Nonparametric Models." https://doi.org/10.1214/23-AOAS1841SUPP

CLARK, T. E., HUBER, F., KOOP, G., MARCELLINO, M. and PFARRHOFER, M., (2023). Tail forecasting with multivariate Bayesian additive regression trees. *Internat. Econom. Rev.* **64** 979–1022. https://doi.org/10.1111/iere.12619

COGLEY, T. and SARGENT, T. J. (2001). Evolving post-World War II US inflation dynamics. *NBER Macroecon. Annu.* **16** 331–373. https://doi.org/10.1086/654451

CRAWFORD, L., FLAXMAN, S. R., RUNCIE, D. E. and WEST, M. (2019). Variable prioritization in nonlinear black box methods: A genetic association case study. *Ann. Appl. Stat.* **13** 958–989. MR3963559 https://doi.org/10.1214/18-AOAS1222

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

FAUST, J. and WRIGHT, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting* (G. Elliott and A. Timmermann, eds.) **2** 2–56. Elsevier, Amsterdam. https://doi.org/10.1016/B978-0-444-53683-9.00001-3

FRÜHWIRTH-SCHNATTER, S. and MALSINER-WALLI, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv. Data Anal. Classif.* **13** 33–64. MR3935190 https://doi.org/10.1007/s11634-018-0329-y

Key words and phrases. Nonparametric regression, Gaussian process, Dirichlet process mixture, inflation forecasting.

GEORGE, E., LAUD, P., LOGAN, B., MCCULLOCH, R. and SPARAPANI, R. (2018). Fully nonparametric Bayesian additive regression trees. arXiv preprint, Available at arXiv:1807.00068.

GIACOMINI, R. and KOMUNJER, I. (2005). Evaluation and combination of conditional quantile forecasts. *J. Bus. Econom. Statist.* **23** 416–431. MR2206011 https://doi.org/10.1198/073500105000000018

GOULET COULOMBE, P., LEROUX, M., STEVANOVIC, D. and SURPRENANT, S. (2022). How is machine learning useful for macroeconomic forecasting? *J. Appl. Econometrics* **37** 920–964. MR4470810 https://doi.org/10.1002/jae.2910

GOULET COULOMBE, P. (2020). The macroeconomy as a random forest. Available at arXiv:2006.12724. https://doi.org/10.48550/arXiv.2006.12724

GOULET COULOMBE, P. (2022). A neural Phillips curve and a deep output gap. arXiv preprint. Available at arXiv:2202.04146. https://doi.org/10.48550/arXiv.2202.04146

GOULET COULOMBE, P., MARCELLINO, M. and STEVANOVIC, D. (2021). Can machine learning catch the COVID-19 recession? *Natl. Inst. Econ. Rev.* **256** 71–109. https://doi.org/10.1017/nie.2021.10

HAUZENBERGER, N., HUBER, F., MARCELLINO, M. and PETZ, N. (2021). Gaussian process Vector Autoregressions and macroeconomic uncertainty. Manuscript. https://doi.org/10.48550/arXiv.2112.01995

HUBER, F. and KOOP, G. (2023). Subspace shrinkage in conjugate Bayesian vector autoregressions. *J. Appl. Econometrics* **38** 556–576. MR4596791

JENSEN, M. J. and MAHEU, J. M. (2010). Bayesian semiparametric stochastic volatility modeling. *J. Econometrics* **157** 306–316. MR2661603 https://doi.org/10.1016/j.jeconom.2010.01.014

JENSEN, M. J. and MAHEU, J. M. (2014). Estimating a semiparametric asymmetric stochastic volatility model with a Dirichlet process mixture. *J. Econometrics* **178** 523–538. MR3132449 https://doi.org/10.1016/j.jeconom.2013.08.018

JOCHMANN, M. (2015). Modeling U.S. inflation dynamics: A Bayesian nonparametric approach. *Econometric Rev.* **34** 537–558. MR3275800 https://doi.org/10.1080/07474938.2013.806199

LEE, J., BAHRI, Y., NOVAK, R., SCHOENHOLZ, S., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2017). Deep neural nets as Gaussian processes. Available at arXiv:1711.00165.

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

LOPEZ-SALIDO, D. and LORIA, F. (2020). Inflation at risk. In *Finance and Economics Discussion Series* 2020-013, *Board of Governors of the Federal Reserve System*. https://doi.org/10.17016/FEDS.2020.013

MASINI, R. P., MEDEIROS, M. C. and MENDES, E. F. (2023). Machine learning advances for time series forecasting. *J. Econ. Surv.* **37** 76–111. https://doi.org/10.1111/joes.12429

MCCRACKEN, M. W. and NG, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Federal Reserve Bank of St. Louis Working Paper 2020-005B. https://doi.org/10.20955/wp.2020.005.

MEDEIROS, M. C., VASCONCELOS, G. F. R., VEIGA, Á. and ZILBERMAN, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *J. Bus. Econom. Statist.* **39** 98–119. MR4187178 https://doi.org/10.1080/07350015.2019.1637745

NAKAMURA, E. (2005). Inflation forecasting using a neural network. *Econom. Lett.* **86** 373–378. MR2124422 https://doi.org/10.1016/j.econlet.2004.09.003

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

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

SHIN, M., BHATTACHARYA, A. and JOHNSON, V. E. (2020). Functional horseshoe priors for subspace shrinkage. *J. Amer. Statist. Assoc.* **115** 1784–1797. MR4189757 https://doi.org/10.1080/01621459.2019.1654875

SIMS, C. A. (2001). Evolving post-World War II US inflation dynamics: Comment. *NBER Macroecon. Annu.* **16** 373–379. https://doi.org/10.1086/654452

STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257 https://doi.org/10.1198/073500102317351921

STOCK, J. H. and WATSON, M. W. (2007). Why has U.S. inflation become harder to forecast? *J. Money Credit Bank.* **39** 3–33. https://doi.org/10.1111/j.1538-4616.2007.00014.x

STOCK, J. H. and WATSON, M. W. (2010). Modeling inflation after the crisis. Working Paper No. 16488, National Bureau of Economic Research. https://doi.org/10.3386/w16488

WOODY, S., CARVALHO, C. M. and MURRAY, J. S. (2021). Model interpretation through lower-dimensional posterior summarization. *J. Comput. Graph. Statist.* **30** 144–161. MR4235972 https://doi.org/10.1080/10618600.2020.1796684

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1445–1470
https://doi.org/10.1214/23-AOAS1842

ANALYZING CROSS-TALK BETWEEN SUPERIMPOSED SIGNALS: VECTOR NORM DEPENDENT HIDDEN MARKOV MODELS AND APPLICATIONS TO ION CHANNELS

BY LAURA JULA VANEGAS[1,a], BENJAMIN ELTZNER[2,c], DANIEL RUDOLF[3,d], MIROSLAV DURA[4,e], STEPHAN E. LEHNART[5,f] AND AXEL MUNK[1,b]

[1]*Institute for Mathematical Stochastics, Georg-August-Universität Göttingen,* [a]*ljulava@mathematik.uni-goettingen.de,* [b]*munk@math.uni-goettingen.de*

[2]*Department for Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences,* [c]*benjamin.eltzner@mpinat.mpg.de*

[3]*Faculty of Computer Science and Mathematics, Universität Passau,* [d]*daniel.rudolf@uni-passau.de*

[4]*Cellular Biophysics and Translational Cardiology Section, Heart Research Center Göttingen, Department of Cardiology & Pneumology, University Medical Center Göttingen,* [e]*miroslav.dura@med.uni-goettingen.de*

[5]*DZHK (German Centre for Cardiovascular Research),* [f]*slehnart@med.uni-goettingen.de*

We propose and investigate a hidden Markov model (HMM) for the analysis of dependent, aggregated, superimposed two-state signal recordings. A major motivation for this work is that often these signals cannot be observed individually but only their superposition. Among others, such models are in high demand for the understanding of cross-talk between ion channels, where each single channel cannot be measured separately. As an essential building block, we introduce a parameterized vector norm dependent Markov chain model and characterize it in terms of permutation invariance as well as conditional independence. This building block leads to a hidden Markov chain sum process which can be used for analyzing the dependence structure of superimposed two-state signal observations within an HMM. Notably, the model parameters of the vector norm dependent Markov chain are uniquely determined by the parameters of the sum process and are, therefore, identifiable. We provide algorithms to estimate the parameters, discuss model selection and apply our methodology to real-world ion channel data from the heart muscle, where we show competitive gating.

REFERENCES

BALL, F., MILNE, R. K., TAME, I. D. and YEO, G. F. (1997). Superposition of interacting aggregated continuous-time Markov chains. *Adv. in Appl. Probab.* **29** 56–91. MR1432931 https://doi.org/10.2307/1427861

BALL, F. G. and RICE, J. A. (1992). Stochastic models for ion channels: Introduction and bibliography. *Math. Biosci.* **112** 189–206. https://doi.org/10.1016/0025-5564(92)90023-p

BARTSCH, A., LLABRÉS, S., PEIN, F., KATTNER, C., SCHÖN, M., DIEHN, M., TANABE, M., MUNK, A., ZACHARIAE, U. et al. (2019). High-resolution experimental and computational electrophysiology reveals weak β-lactam binding events in the porin PorB. *Sci. Rep.* **9** 1264. https://doi.org/10.1038/s41598-018-37066-9

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

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. MR0287613 https://doi.org/10.1214/aoms/1177697196

BECKER, J. D., HONERKAMP, J., HIRSCH, J., FRÖBE, U., SCHLATTER, E. and GREGER, R. (1994). Analysing ion channels with hidden Markov models. *Pflügers Arch.* **426** 328–332.

BEHR, M., HOLMES, C. and MUNK, A. (2018). Multiscale blind source separation. *Ann. Statist.* **46** 711–744. MR3782382 https://doi.org/10.1214/17-AOS1565

Key words and phrases. Hidden Markov models, vector norm dependency, permutation invariance, lumping property, aggregated data, crosstalk, ion channels.

BICKEL, P. J., RITOV, Y. and RYDÉN, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614–1635. MR1647705 https://doi.org/10.1214/aos/1024691255

BIELECKI, T. R., JAKUBOWSKI, J. and NIEWĘGŁOWSKI, M. (2013). Intricacies of dependence between components of multivariate Markov chains: Weak Markov consistency and weak Markov copulae. *Electron. J. Probab.* **18** 45. MR3040555 https://doi.org/10.1214/EJP.v18-2238

BRAND, M., OLIVER, N. and PENTLAND, A. (1997). Coupled hidden Markov models for complex action recognition. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 994–999.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer, New York. MR2159833

CELEUX, G. and DURAND, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.* **23** 541–564. MR2448181 https://doi.org/10.1007/s00180-007-0097-1

CHEN, C., LIANG, J., ZHAO, H., HU, H. and TIAN, J. (2009). Factorial HMM and parallel HMM for gait recognition. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **39** 114–123.

CHEN, W., WASSERSTROM, J. A. and SHIFERAW, Y. (2009). Role of coupled gating between cardiac ryanodine receptors in the genesis of triggered arrhythmias. *Am. J. Physiol. Heart Circ. Physiol.* **297** H171–H180. https://doi.org/10.1152/ajpheart.00098.2009

CHEN, Y., SHEN, K., SHAN, S.-O. and KOU, S. C. (2016). Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *J. Amer. Statist. Assoc.* **111** 951–966. MR3561922 https://doi.org/10.1080/01621459.2016.1140050

CHUNG, S.-H., ANDERSON, O. S. and KRISHNAMURTHY, V. V., eds. (2007) *Biological Membrane Ion Channels*: *Dynamics*, *Structure*, *and Applications*. *Biological and Medical Physics*, *Biomedical Engineering*. Springer, New York.

CHUNG, S. H. and KENNEDY, R. A. (1996). Coupled Markov chain model: Characterization of membrane channel currents with multiple conductance sublevels as partially coupled elementary pores. *Math. Biosci.* **133** 111–137. https://doi.org/10.1016/0025-5564(95)00084-4

CSISZÁR, I. and SHIELDS, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28** 1601–1619. MR1835033 https://doi.org/10.1214/aos/1015957472

DABROWSKI, A. R. and MCDONALD, D. (1992). Statistical analysis of multiple ion channel data. *Ann. Statist.* **20** 1180–1202. MR1186246 https://doi.org/10.1214/aos/1176348765

DE GUNST, M. C. M., KÜNSCH, H. R. and SCHOUTEN, J. G. (2001). Statistical analysis of ion channel data using hidden Markov models with correlated state-dependent noise and filtering. *J. Amer. Statist. Assoc.* **96** 805–815. MR1946357 https://doi.org/10.1198/016214501753208519

DIEHN, M., MUNK, A. and RUDOLF, D. (2019). Maximum likelihood estimation in hidden Markov models with inhomogeneous noise. *ESAIM Probab. Stat.* **23** 492–523. MR3989601 https://doi.org/10.1051/ps/2018017

FINE, S., SINGER, Y. and TISHBY, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Mach. Learn.* **32** 41–62.

FREDKIN, D. R. and RICE, J. A. (1991). On the superposition of currents from ion channels. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **334** 347–356. https://doi.org/10.1098/rstb.1991.0121

GALES, M. and YOUNG, S. (2008). *The Application of Hidden Markov Models in Speech Recognition*. Now Publishers, Hanover.

GASSIAT, E. and BOUCHERON, S. (2003). Optimal error exponents in hidden Markov models order estimation. *IEEE Trans. Inf. Theory* **49** 964–980. MR1984482 https://doi.org/10.1109/TIT.2003.809574

GHAHRAMANI, Z. and JORDAN, M. I. (1997). Factorial hidden Markov models. *Mach. Learn.* **29** 245–273.

GNANASAMBANDAM, R., NIELSEN, M. S., NICOLAI, C., SACHS, F., HOFGAARD, J. P. and DREYER, J. K. (2017). Unsupervised idealization of ion channel recordings by minimum description length: Application to human PIEZO1-channels. *Front. Neuroinform.* **11** 31. https://doi.org/10.3389/fninf.2017.00031

GOTTSCHAU, A. (1992). Exchangeability in multivariate Markov chain models. *Biometrics* **48** 751–763.

GUAN, X., RAICH, R. and WONG, W.-K. (2016). Efficient multi-instance learning for activity recognition from time series data using an auto-regressive hidden Markov model. In *Proceedings of the* 33*rd International Conference on Machine Learning—Volume* 48. *ICML'*16 2330–2339. JMLR.org, New York, NY, USA.

JULA VANEGAS, L., BEHR, M. and MUNK, A. (2022). Multiscale quantile segmentation. *J. Amer. Statist. Assoc.* **117** 1384–1397. MR4480719 https://doi.org/10.1080/01621459.2020.1859380

KELESHIAN, A. M., EDESON, R. O., LIU, G.-J. and MADSEN, B. W. (2000). Evidence for cooperativity between nicotinic acetylcholine receptors in patch clamp records. *Biophys. J.* **78** 1–12.

KEMENY, J. G. and SNELL, J. L. (1976). *Finite Markov Chains*: *With a New Appendix "Generalization of a Fundamental Matrix"*. *Undergraduate Texts in Mathematics*. Springer, New York.

KHAN, R. N., MARTINAC, B., MADSEN, B. W., MILNE, R. K., YEO, G. F. and EDESON, R. O. (2005). Hidden Markov analysis of mechanosensitive ion channel gating. *Math. Biosci.* **193** 139–158. MR2123740 https://doi.org/10.1016/j.mbs.2004.07.007

KLEIN, S., TIMMER, J. and HONERKAMP, J. (1997). Analysis of multichannel patch clamp recordings by hidden Markov models. *Biometrics* **53** 870–884.

KROGH, A., LARSSON, B., VON HEIJNE, G. and SONNHAMMER, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305** 567–580.

LAVER, D. R., O'NEILL, E. R. and LAMB, G. D. (2004). Luminal Ca2+-regulated Mg2+ inhibition of skeletal RyRs reconstituted as isolated channels or coupled clusters. *J. Gen. Physiol.* **124** 741–758. https://doi.org/10.1085/jgp.200409092

LEHÉRICY, L. (2019). Consistent order estimation for nonparametric hidden Markov models. *Bernoulli* **25** 464–498. MR3892326 https://doi.org/10.3150/17-bej993

MANOGARAN, G., VIJAYAKUMAR, V., VARATHARAJAN, R., MALARVIZHI KUMAR, P., SUNDARASEKAR, R. and HSU, C.-H. (2018). Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. *Wirel. Pers. Commun.* **102** 2099–2116.

MARI, J. F., HATON, J. P. and KRIOUILE, A. (1997). Automatic word recognition based on second-order hidden Markov models. *IEEE Trans. Speech Audio Process.* **5** 22–25.

MARX, S. O., GABURJÁKOVÁ, J., GABURJAKOVA, M., HENRIKSON, C. A., ONDRIAS, K. and MARKS, A. R. (2001). Coupled gating between cardiac calcium release channels (ryanodine receptors). *Circ. Res.* **88** 1151–1158.

MIRAMS, G. R., CUI, Y., SHER, A., FINK, M., COOPER, J., HEATH, B. M., MCMAHON, N. C., GAVAGHAN, D. J. and NOBLE, D. (2011). Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk. *Cardiovasc. Res.* **91** 53–61. https://doi.org/10.1093/cvr/cvr044

NEUKIRCH, M., RUDOLF, D., GARCIA, X. and GALIANA, S. (2019). Amplitude-phase decomposition of the magnetotelluric impedance tensor. *Geophysics* **84** E301–E310.

PEIN, F., ELTZNER, B. and MUNK, A. (2021). Analysis of patchclamp recordings: Model-free multiscale methods and software. *Eur. Biophys. J.* **50** 187–209. https://doi.org/10.1007/s00249-021-01506-8

PEIN, F., TECUAPETLA-GOMEZ, I., SCHUTTE, O. M., STEINEM, C. and MUNK, A. (2018). Fully automatic multiresolution idealization for filtered ion channel recordings: Flickering event detection. *IEEE Trans. Nanobiosci.* **17** 300–320. https://doi.org/10.1109/TNB.2018.2845126

PERKEL, J. M. (2010). High-throughput ion channel screening: A "patch"-work solution. *BioTechniques* **48** 25–29. https://doi.org/10.2144/000113339

PORTA, M., DIAZ-SYLVESTER, P. L., NEUMANN, J. T., ESCOBAR, A. L., FLEISCHER, S. and COPELLO, J. A. (2012). Coupled gating of skeletal muscle ryanodine receptors is modulated by Ca2+, Mg2+, and ATP. *Am. J. Physiol., Cell Physiol.* **303** C682–C697. https://doi.org/10.1152/ajpcell.00150.2012

SAKMANN, B. and NEHER, E., eds. (1995) *Single-Channel Recording*, 2nd ed. Springer, New York.

SALVAGE, S. C., GALLANT, E. M., BEARD, N. A., AHMAD, S., VALLI, H., FRASER, J. A., HUANG, C. L. H. and DULHUNTY, A. F. (2019). Ion channel gating in cardiac ryanodine receptors from the arrhythmic RyR2-P2328S mouse. *J. Cell Sci.* **132**.

SCHMIDT-HIEBER, J., SCHNEIDER, L. F., STAUDT, T., KRAJINA, A., ASPELMEIER, T. and MUNK, A. (2021). Posterior analysis of n in the binomial (n, p) problem with both parameters unknown—with applications to quantitative nanoscopy. *Ann. Statist.* **49** 3534–3558. MR4352540 https://doi.org/10.1214/21-aos2096

SHERLOCK, C., XIFARA, T., TELFER, S. and BEGON, M. (2013). A coupled hidden Markov model for disease interactions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 609–627. MR3083914 https://doi.org/10.1111/rssc.12015

SIEKMANN, I., FACKRELL, M., CRAMPIN, E. J. and TAYLOR, P. (2016). Modelling modal gating of ion channels with hierarchical Markov models. *Proc. R. Soc. A, Math. Phys. Eng. Sci.* **472** 20160122. MR3551029 https://doi.org/10.1098/rspa.2016.0122

SIN, B. and KIM, J. H. (1995). Nonstationary hidden Markov model. *Signal Process.* **46** 31–46.

STAUDT, T., ASPELMEIER, T., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2020). Statistical molecule counting in super-resolution fluorescence microscopy: Towards quantitative nanoscopy. *Statist. Sci.* **35** 92–111. MR4071360 https://doi.org/10.1214/19-STS753

TAUR, Y. and FRISHMAN, W. (2005). The cardiac ryanodine receptor (RyR2) and its role in heart disease. *Cardiol. Rev.* **13** 142–146.

TOULOUPOU, P., FINKENSTÄDT, B. and SPENCER, S. E. F. (2020). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *J. Comput. Graph. Statist.* **29** 238–249. MR4116038 https://doi.org/10.1080/10618600.2019.1654880

VAN DER KAMP, W. S. and OSGOOD, N. D. (2017). Multivariate hidden Markov models for personal smartphone sensor data: Time series analysis. In 2017 *IEEE Int. Conf. Healthc. Inform.* 179–188. https://doi.org/10.1109/ICHI.2017.84

JULA VANEGAS, L., ELTZNER, B., RUDOLF, D., DURA, M., LEHNART, S. E. and MUNK, A. (2024). Supplement to "Analyzing cross-talk between superimposed signals: Vector norm dependent hidden Markov models and applications to ion channels." https://doi.org/10.1214/23-AOAS1842SUPP

VENKATARAMANAN, L. and SIGWORTH, F. J. (2002). Applying hidden Markov models to the analysis of single ion channel activity. *Biophys. J.* **82** 1930–1942.

WALKER, M. A., KOHL, T., LEHNART, S. E., GREENSTEIN, J. L., LEDERER, W. J. and WINSLOW, R. L. (2015). On the adjacency matrix of RyR2 cluster structures. *PLoS Comput. Biol.* **11** 1–21. https://doi.org/10.1371/journal.pcbi.1004521

WALKER, M. A., WILLIAMS, G. S. B., KOHL, T., LEHNART, S. E., JAFRI, M. S., GREENSTEIN, J. L., LEDERER, W. J. and WINSLOW, R. L. (2014). Superresolution modeling of calcium release in the heart. *Biophys. J.* **107** 3018–3029. https://doi.org/10.1016/j.bpj.2014.11.003

WESTHEAD, D. R. and VIJAYABASKAR, M. (2017). *Hidden Markov Models*: *Methods and Protocols*. Springer, Berlin.

WILLIAMS, A. J., THOMAS, N. L. and GEORGE, C. H. (2018). The ryanodine receptor: Advances in structure and organization. *Curr. Opin. Physiol.* **1** 1–6.

YEO, G. F., EDESON, R. O., MILNE, R. K. and MADSEN, B. W. (1989). Superposition properties of independent ion channels. *Proc. R. Soc. Lond., B Biol. Sci.* **238** 155–170. https://doi.org/10.1098/rspb.1989.0073

YONEKURA, S., BESKOS, A. and SINGH, S. S. (2021). Asymptotic analysis of model selection criteria for general hidden Markov models. *Stochastic Process. Appl.* **132** 164–191. MR4179109 https://doi.org/10.1016/j.spa.2020.10.006

ZHANG, Y. and KASSAM, S. A. (2001). Blind separation and equalization using fractional sampling of digital communications signals. *Signal Process.* **81** 2591–2608.

ZUCCHINI, W., MACDONALD, I. L. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series*: *An Introduction Using R*, 2nd ed. *Monographs on Statistics and Applied Probability* **150**. CRC Press, Boca Raton, FL. MR3618333

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1471–1489
https://doi.org/10.1214/23-AOAS1843

FLEXIBLE INSTRUMENTAL VARIABLE MODELS WITH BAYESIAN ADDITIVE REGRESSION TREES

BY CHARLES SPANBAUER[a] AND WEI PAN[b]

Division of Biostatistics, University of Minnesota, [a]*spanb008@umn.edu,* [b]*panxx014@umn.edu*

Methods utilizing instrumental variables have been a fundamental statistical approach to causal estimation in the presence of unmeasured confounding, usually occurring in nonrandomized observational data common to fields such as economics and public health. However, such methods traditionally make constricting linearity and additivity assumptions that are inapplicable to the complex modeling challenges of today. The growing body of observational data being collected may benefit from flexible regression modeling while also retaining the ability to control for confounding using instrumental variables. Therefore, this article presents a flexible instrumental variable regression model based on Bayesian regression tree ensembles to estimate the causal exposure-outcome relationship, including interactions with covariates, in the presence of confounding. One exciting application of this method is to use genetic variants as instruments, known as Mendelian randomization. We present our flexible Bayesian instrumental variable regression tree method with an example from the UK Biobank where body mass index is related to blood pressure using genetic variants as the instruments. Body mass index is one factor that is hypothesized to have a nonlinear relationship with cardiovascular risk factors, such as blood pressure, while interacting with age. Heterogeneity in patient characteristics, such as age, could be clinically interesting from a precision medicine perspective where individualized treatment is emphasized.

REFERENCES

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

BARGAGLI-STOFFI, F. J., DE WITTE, K. and GNECCO, G. (2022). Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach. *Ann. Appl. Stat.* **16** 1986–2009. MR4455908 https://doi.org/10.1214/21-aoas1579

BENNETT, A., KALLUS, N. and SCHNABEL, T. (2019). Deep generalized method of moments for instrumental variable analysis. *Adv. Neural Inf. Process. Syst.* **32**.

BOTEV, J., ÉGERT, B. and JAWADI, F. (2019). The nonlinear relationship between economic growth and financial development: Evidence from developing, emerging and advanced economies. *Int. Econ.* **160** 3–13.

BURGESS, S., DAVIES, N. M., THOMPSON, S. G. and EPIC-INTERACT CONSORTIUM (2014). Instrumental variable analysis with a nonlinear exposure-outcome relationship. *Epidemiology* **25** 877–885. https://doi.org/10.1097/EDE.0000000000000161

BURKE, M., HSIANG, S. M. and MIGUEL, E. (2015). Global non-linear effect of temperature on economic production. *Nature* **527** 235–239. https://doi.org/10.1038/nature15725

CARD, D. (1999). The causal effect of education on earnings. *Handb. Labor Econ.* **3** 1801–1863.

CHETVERIKOV, D. and WILHELM, D. (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica* **85** 1303–1320. MR3681772 https://doi.org/10.3982/ECTA13639

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

DESHPANDE, S. K., BAI, R., BALOCCHI, C., STARLING, J. E. and WEISS, J. (2022). VCBART: Bayesian trees for varying coefficients. arXiv preprint. Available at arXiv:2003.06416.

DZAU, V. J. and GINSBURG, G. S. (2016). Realizing the full potential of precision medicine in health and health care. *JAMA* **316** 1659–1660. https://doi.org/10.1001/jama.2016.14117

Key words and phrases. Causality, genetics, instrumental variables, machine learning, Mendelian randomization.

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

FAWCETT, K. A. and BARROSO, I. (2010). The genetics of obesity: FTO leads the way. *Trends Genet.* **26** 266–274.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67. With discussion and a rejoinder by the author. MR1091842 https://doi.org/10.1214/aos/1176347963

GRINBERG, N. F. and WALLACE, C. (2021). Multi-tissue transcriptome-wide association studies. *Genet. Epidemiol.* **45** 324–337.

GUO, Z. and SMALL, D. S. (2016). Control function instrumental variable estimation of nonlinear causal effect models. *J. Mach. Learn. Res.* **17** Paper No. 100, 35. MR3543506

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. Includes comments and discussions by 25 discussants and a rejoinder by the authors. MR4154846 https://doi.org/10.1214/19-BA1195

HALL, P. and HOROWITZ, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107 https://doi.org/10.1214/009053605000000714

HARTFORD, J., LEWIS, G., LEYTON-BROWN, K. and TADDY, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning* 1414–1423. PMLR.

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

HOROWITZ, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica* **79** 347–394. MR2809374 https://doi.org/10.3982/ECTA8662

JIANG, X., HOLMES, C. and MCVEAN, G. (2021). The impact of age on genetic risk for common diseases. *PLoS Genet.* **17** e1009723. https://doi.org/10.1371/journal.pgen.1009723

JOHNSON, M., CAO, J. and KANG, H. (2022). Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment. *Ann. Appl. Stat.* **16** 1111–1129. MR4438826 https://doi.org/10.1214/21-aoas1535

KIEL, L. D. (2000). The evolution of nonlinear dynamics in political science and public administration: Methods, modeling and momentum. *Discrete Dyn. Nat. Soc.* **5** 265–279.

LANDI, F., CALVANI, R., PICCA, A., TOSATO, M., MARTONE, A. M. ORTOLANI, E., SISTO, A., D'ANGELO, E., SERAFINI, E. et al. (2018). Body mass index is strongly associated with hypertension: Results from the longevity check-up 7+ study. *Nutrients* **10**. 1976.

LI, B. and RITCHIE, M. D. (2021). From GWAS to gene: Transcriptome-wide association studies and other methods to functionally understand GWAS discoveries. *Front. Genet.* **12** 713230. https://doi.org/10.3389/fgene.2021.713230

LI, H., MA, J., ZHENG, D., LI, X., GUO, X., WANG, J. and SU, P. (2021). Sex differences in the non-linear association between BMI and LDL cholesterol in middle-aged and older adults: Findings from two nationally representative surveys in China. *Lipids Health Dis.* **20** 1–12.

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

LINK, W. A. and EATON, M. J. (2012). On thinning of chains in MCMC. *Methods Ecol. Evol.* **3** 112–115.

LIPSITZ, K. and PADILLA, J. (2021). The nonlinear effects of political advertising. *J. Polit. Mark.* 1–14.

LOGAN, B. R., SPARAPANI, R., MCCULLOCH, R. E. and LAUD, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Stat. Methods Med. Res.* **28** 1079–1093. MR3934636 https://doi.org/10.1177/0962280217746191

LOPES, H. F. and POLSON, N. G. (2014). Bayesian instrumental variables: Priors and likelihoods. *Econometric Rev.* **33** 100–121. MR3170842 https://doi.org/10.1080/07474938.2013.807146

LOUSDAL, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology* **15** 1–7.

MCCULLOCH, R. E., SPARAPANI, R. A., LOGAN, B. R. and LAUD, P. W. (2021). Causal inference with the instrumental variable approach and Bayesian nonparametric machine learning. arXiv preprint. Available at arXiv:2102.01199.

MURRAY, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *J. Amer. Statist. Assoc.* **116** 756–769. MR4270022 https://doi.org/10.1080/01621459.2020.1813587

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257 https://doi.org/10.1111/1468-0262.00459

OKORO, P. C., SCHUBERT, R., GUO, X., JOHNSON, W. C. ROTTER, J. I., HOESCHELE, I., LIU, Y., IM, H. K., LUKE, A. et al. (2021). Transcriptome prediction performance across machine learning models and diverse ancestries. *Hum. Genet. Genomics Adv.* **2** 100019.

PETER, R. S., MAYER, B., CONCIN, H. and NAGEL, G. (2015). The effect of age on the shape of the BMI–mortality relation and BMI associated with minimum all-cause mortality in a large Austrian cohort. *Int. J. Obes.* **39** 530–534.

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

ROSSI, P. E., ALLENBY, G. M. and MCCULLOCH, R. (2005). *Bayesian Statistics and Marketing*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2193403 https://doi.org/10.1002/0470863692

SCARNECIU, C. C., SANGEORZAN, L., RUS, H., SCARNECIU, V. D., VARCIU, M. S., ANDREESCU, O. and SCARNECIU, I. (2017). Comparison of linear and non-linear regression analysis to determine pulmonary pressure in hyperthyroidism. *Pak. J. Med. Sci.* **33** 111–120. https://doi.org/10.12669/pjms.331.11046

SPANBAUER, C., PAN, W. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2022). Sparse prediction informed by genetic annotations using the logit normal prior for Bayesian regression tree ensembles. *Genet. Epidemiol.*

SPANBAUER, C. and PAN, W. (2024). Supplement to "Flexible instrumental variable models with Bayesian additive regression trees." https://doi.org/10.1214/23-AOAS1843SUPPA, https://doi.org/10.1214/23-AOAS1843SUPPB

SPANBAUER, C. and SPARAPANI, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. *Stat. Med.* **40** 2665–2691. MR4255774 https://doi.org/10.1002/sim.8924

SPARAPANI, R., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2020a). Nonparametric competing risks analysis using Bayesian additive regression trees. *Stat. Methods Med. Res.* **29** 57–77. MR4055122 https://doi.org/10.1177/0962280218822140

SPARAPANI, R., SPANBAUER, C. and MCCULLOCH, R. E. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *J. Stat. Softw.* **97** 1–66. https://doi.org/10.18637/jss.v097.i01

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

SPARAPANI, R. A., REIN, L. E., TARIMA, S. S., JACKSON, T. A. and MEURER, J. R. (2020b). Non-parametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes. *Biostatistics* **21** 69–85. MR4043846 https://doi.org/10.1093/biostatistics/kxy032

STOCK, J. H. and TREBBI, F. (2003). Retrospectives: Who invented instrumental variable regression? *J. Econ. Perspect.* **17** 177–194.

TAN, Y. V., FLANNAGAN, C. A. C. and ELLIOTT, M. R. (2018). Predicting human-driving behavior to help driverless vehicles drive: Random intercept Bayesian additive regression trees. *Stat. Interface* **11** 557–572. MR3858513 https://doi.org/10.4310/SII.2018.v11.n4.a1

TUU, H. H. and OLSEN, S. O. (2010). Nonlinear effects between satisfaction and loyalty: An empirical study of different conceptual relationships. *J. Target. Meas. Anal. Mark.* **18** 239–251.

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

WIESENFARTH, M., HISGEN, C. M., KNEIB, T. and CADARSO-SUAREZ, C. (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and Dirichlet process mixtures. *J. Bus. Econom. Statist.* **32** 468–482. MR3238599 https://doi.org/10.1080/07350015.2014.907092

XUE, H., PAN, W. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2020). Some statistical consideration in transcriptome-wide association studies. *Genet. Epidemiol.* **44** 221–232.

YIN, D., BOND, S. D. and ZHANG, H. (2017). Keep your cool or let it out: Nonlinear effects of expressed arousal on perceptions of consumer reviews. *J. Mark. Res.* **54** 447–463.

ZACCARDI, F., DHALWANI, N. N., PAPAMARGARITIS, D., WEBB, D. R., MURPHY, G. J., DAVIES, M. J. and KHUNTI, K. (2017). Nonlinear association of BMI with all-cause and cardiovascular mortality in type 2 diabetes mellitus: A systematic review and meta-analysis of 414,587 participants in prospective studies. *Diabetologia* **60** 240–248.

ZELDOW, B., LO RE III, V. 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

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1490–1505
https://doi.org/10.1214/23-AOAS1844

PENALIZED JOINT MODELS OF HIGH-DIMENSIONAL LONGITUDINAL BIOMARKERS AND A SURVIVAL OUTCOME

BY JIEHUAN SUN[a] AND SANJIB BASU[b]

Division of Epidemiology and Biostatistics, University of Illinois Chicago, [a]*jiehuan@uic.edu,* [b]*sbasu@uic.edu*

High-dimensional biomarkers, such as gene expression profiles, are often collected longitudinally to monitor disease progression in clinical studies, where the primary endpoint of interest is often a survival outcome. It is of great interest to study the associations between high-dimensional longitudinal biomarkers and the survival outcome as well as to identify biomarkers related to the survival outcome. Joint models, which have been extensively studied in the past decades, are commonly used to study the associations between longitudinal biomarkers and the survival outcome. However, existing joint models only consider one or a few longitudinal biomarkers and cannot deal with high-dimensional longitudinal biomarkers. In this paper we propose a novel penalized joint model that can handle high-dimensional longitudinal biomarkers. Specifically, we impose an adaptive lasso penalty on the parameters for the effects of the longitudinal biomarkers on the survival outcome, which allows for variable selection. We also develop a computationally efficient algorithm for model estimation based on the Gaussian variational approximation method, which can be implemented using the HDJM package in R. Furthermore, based on the penalized joint model, we propose a two-stage selection procedure that can reduce the estimation bias, due to the penalization, and allows for inference. We conduct extensive simulation studies to evaluate the performance of our proposed method. The performance of our proposed method is further demonstrated on a longitudinal gene expression dataset of patients with idiopathic pulmonary fibrosis.

REFERENCES

AHANGARI, F., BECKER, C., FOSTER, D. G., CHIOCCIOLI, M., NELSON, M., BEKE, K., WANG, X., JUSTET, A., ADAMS, T. et al. (2022). Saracatinib, a selective Src kinase inhibitor, blocks fibrotic responses in preclinical models of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **206** 1463–1479.

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

BARRETT, J. and SU, L. (2017). Dynamic predictions using flexible joint models of longitudinal and time-to-event data. *Stat. Med.* **36** 1447–1460. MR3631971 https://doi.org/10.1002/sim.7209

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

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

CHEN, Y. and WANG, Y. (2017). Variable selection for joint models of multivariate longitudinal measurements and event time data. *Stat. Med.* **36** 3820–3829. MR3713143 https://doi.org/10.1002/sim.7391

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871 https://doi.org/10.1111/rssb.12033

DAS, D., GREGORY, K. and LAHIRI, S. N. (2019). Perturbation bootstrap in adaptive Lasso. *Ann. Statist.* **47** 2080–2116. MR3953445 https://doi.org/10.1214/18-AOS1741

Key words and phrases. Adaptive lasso, high-dimensional longitudinal data, joint models, survival outcome, variational approximation.

DE BRUIJN, N. G. (1981). *Asymptotic Methods in Analysis*, 3rd ed. Dover, New York. MR0671583
DEPIANTO, D. J., CHANDRIANI, S., ABBAS, A. R., JIA, G., N'DIAYE, E. N., CAPLAZI, P., KAUDER, S. E., BISWAS, S., KARNIK, S. K. et al. (2015). Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. *Thorax* **70** 48–56.
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
FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 https://doi.org/10.1214/aos/1015362185
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
FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.
HALL, P., PHAM, T., WAND, M. P. and WANG, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39** 2502–2532. MR2906876 https://doi.org/10.1214/11-AOS908
HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. (2020). Best subset, forward stepwise or Lasso? Analysis and recommendations based on extensive comparisons. *Statist. Sci.* **35** 579–592. MR4175382 https://doi.org/10.1214/19-STS733
HASTIE, T., TIBSHIRANI, R., EISEN, M. B., ALIZADEH, A., LEVY, R., STAUDT, L., CHAN, W. C., BOTSTEIN, D. and BROWN, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* **1** 1–21.
HE, Z., TU, W., WANG, S., FU, H. and YU, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* **71** 178–187. MR3335362 https://doi.org/10.1111/biom.12221
HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.
HERAZO-MAYA, J. D., NOTH, I., DUNCAN, S. R., KIM, S., MA, S.-F., TSENG, G. C., FEINGOLD, E., JUAN-GUARDELA, B. M., RICHARDS, T. J. et al. (2013). Peripheral blood mononuclear cell gene expression profiles predict poor outcome in idiopathic pulmonary fibrosis. *Sci. Transl. Med.* **5** 205ra136–205ra136.
HERAZO-MAYA, J. D., SUN, J., MOLYNEAUX, P. L., LI, Q., VILLALBA, J. A., TZOUVELEKIS, A., LYNN, H., JUAN-GUARDELA, B. M., RISQUEZ, C. et al. (2017). Validation of a 52-gene risk profile for outcome prediction in patients with idiopathic pulmonary fibrosis: An international, multicentre, cohort study. *Lancet Respir. Med.* **5** 857–868.
HSIEH, F., TSENG, Y.-K. and WANG, J.-L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics* **62** 1037–1043. MR2297674 https://doi.org/10.1111/j.1541-0420.2006.00570.x
KERIOUI, M., MERCIER, F., BERTRAND, J., TARDIVON, C., BRUNO, R., GUEDJ, J. and DESMÉE, S. (2020). Bayesian inference using Hamiltonian Monte-Carlo algorithm for nonlinear joint modeling in the context of cancer immunotherapy. *Stat. Med.* **39** 4853–4868. MR4190950 https://doi.org/10.1002/sim.8756
LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371
LEY, B., RYERSON, C. J., VITTINGHOFF, E., RYU, J. H., TOMASSETTI, S., LEE, J. S., POLETTI, V., BUCCIOLI, M., ELICKER, B. M. et al. (2012). A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann. Intern. Med.* **156** 684–691.
LIU, M., SUN, J., HERAZO-MAYA, J. D., KAMINSKI, N. and ZHAO, H. (2019). Joint models for time-to-event data and longitudinal biomarkers of high dimension. *Stat. Biosci.* **11** 614–629.
MA, Z., DAVIS, S. W. and HO, Y.-Y. (2023). Flexible copula model for integrating correlated multi-omics data from single-cell experiments. *Biometrics* **79** 1559–1572. MR4606374
MOLYNEAUX, P. L., WILLIS-OWEN, S. A. G., COX, M. J., JAMES, P., COWMAN, S., LOEBINGER, M., BLANCHARD, A., EDWARDS, L. M., STOCK, C. et al. (2017). Host–microbial interactions in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **195** 1640–1650.
ORMEROD, J. T. and WAND, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Statist.* **21** 2–17. MR2913353 https://doi.org/10.1198/jcgs.2011.09118
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
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
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

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
SINNOTT, J. A. and CAI, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics* **17** 692–707. MR3604274 https://doi.org/10.1093/biostatistics/kxw016
SUN, J. and BASU, S. (2024). Supplement to "Penalized joint models of high-dimensional longitudinal biomarkers and a survival outcome." https://doi.org/10.1214/23-AOAS1844SUPP
SUN, J., HERAZO-MAYA, J. D., MOLYNEAUX, P. L., MAHER, T. M., KAMINSKI, N. and ZHAO, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics* **75** 69–77. MR3953708 https://doi.org/10.1111/biom.12964
TANG, A.-M., ZHAO, X. and TANG, N.-S. (2017). Bayesian variable selection and estimation in semiparametric joint models of multivariate longitudinal and survival data. *Biom. J.* **59** 57–78. MR3593721 https://doi.org/10.1002/bimj.201500070
TAYLOR, J. M. G., PARK, Y., ANKERST, D. P., PROUST-LIMA, C., WILLIAMS, S., KESTIN, L., BAE, K., PICKLES, T. and SANDLER, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* **69** 206–213. MR3058067 https://doi.org/10.1111/j.1541-0420.2012.01823.x
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974
TU, J. and SUN, J. (2023). Gaussian variational approximate inference for joint models of longitudinal biomarkers and a survival outcome. *Stat. Med.* **42** 316–330. MR4537851
VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
WANG, Y. and TAYLOR, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J. Amer. Statist. Assoc.* **96** 895–905. MR1946362 https://doi.org/10.1198/016214501753208591
WU, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statist. Sinica* **22** 271–294. MR2933176 https://doi.org/10.5705/ss.2010.107
XIE, Y., HE, Z., TU, W. and YU, Z. (2020). Variable selection for joint models with time-varying coefficients. *Stat. Methods Med. Res.* **29** 309–322. MR4055139 https://doi.org/10.1177/0962280219873125
XU, J. and ZEGER, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *J. R. Stat. Soc., Ser. C* **50** 375–387. MR1856332 https://doi.org/10.1111/1467-9876.00241
YI, F., TANG, N. and SUN, J. (2022). Simultaneous variable selection and estimation for joint models of longitudinal and failure time data with interval censoring. *Biometrics* **78** 151–164. MR4408577 https://doi.org/10.1111/biom.13387
YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018
ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729
ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026
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
2024, Vol. 18, No. 2, 1506–1518
https://doi.org/10.1214/23-AOAS1846

AS TREATED ANALYSES OF CLUSTER RANDOMIZED TRIALS

BY ARI I. F. FOGELSON[1,a], KIRSTEN E. LANDSIEDEL[2,b], SUZANNE M. DUFAULT[3,c] AND NICHOLAS P. JEWELL[4,d]

[1]*Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine,* [a]*Ari.Fogelson@lshtm.ac.uk*

[2]*Division of Biostatistics, School of Public Health, University of California, Berkeley,* [b]*kirsten_landsiedel@berkeley.edu*

[3]*Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco,* [c]*Suzanne.Dufault@ucsf.edu*

[4]*Department of Medical Statistics, London School of Hygiene and Tropical Medicine,* [d]*Nicholas.Jewell@lshtm.ac.uk*

Test-negative designs have rapidly become an appealing approach to assess disease interventions when randomization is not feasible and specifically used to measure the effectiveness of vaccines in the field (*Vaccine* **31** (2013) 2165–2168). An innovative extension of the test-negative design was recently used to assess the impact of a mosquito intervention where the intervention was applied at a cluster level with cluster assignment chosen at random, the AWED (applying *Wolbachia* to eliminate dengue) trial. The primary analysis reported was intention-to-treat (ITT) (*Trials* **19** (2018) 302; *N. Engl. J. Med.* **384** (2021) 2177–2186). However, the level of uptake of the intervention on mosquitoes was routinely captured in all clusters over time, and, furthermore, participants' mobility across clusters was measured in the time immediately preceding the onset of symptoms (whether test-positive or test-negative). Combinations of these measurements provide proxies for the true exposure to the intervention, thereby permitting an "as treated" assessment. We consider the use of marginal generalized estimating equations (GEE) and conditional generalized inear mixed models (GLMM) to estimate as treated efficacy, contrasting both with the ITT. We illustrate the strengths and challenges of these methods in the context of the AWED trial, highlighting several ways that common approaches to analysis of clustered data can yield incorrect results that can in turn be obscured and compounded by limitations in routine software. In addition, we estimate a greater level of intervention efficacy than shown in the ITT analysis.

REFERENCES

AGBLA, S. C., DE STAVOLA, B. and DIAZORDAZ, K. (2020). Estimating cluster-level local average treatment effects in cluster randomised trials with non-adherence. *Stat. Methods Med. Res.* **29** 911–933. MR4078257 https://doi.org/10.1177/0962280219849613

AGBLA, S. C. and DIAZORDAZ, K. (2018). Reporting non-adherence in cluster randomised trials: A systematic review. *Clin. Trials* **15** 294–304. https://doi.org/10.1177/1740774518761666

ANDERS, K. L., CUTCHER, Z., KLEINSCHMIDT, I., DONNELLY, C. A., FERGUSON, N. M., INDRIANI, C., RYAN, P. A., O'NEILL, S. L., JEWELL, N. P. et al. (2018a). Cluster-randomized test-negative design trials: A novel and efficient method to assess the efficacy of community-level Dengue interventions. *Amer. J. Epidemiol.* **187** 2021–2028. https://doi.org/10.1093/AJE/KWY099

ANDERS, K. L., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANDARI, B., JEWELL, N. P., DUFAULT, S. M., RYAN, P. A. et al. (2020). Update to the AWED (Applying Wolbachia to Eliminate Dengue) trial study protocol: A cluster randomised controlled trial in Yogyakarta, Indonesia. *Trials* **21** 429. https://doi.org/10.1186/s13063-020-04367-2

ANDERS, K. L., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANDARI, B., JEWELL, N. P., RANCES, E., O'NEILL, S. L. et al. (2018b). The AWED trial (Applying Wolbachia to Eliminate Dengue) to assess the efficacy of Wolbachia-infected mosquito deployments to reduce dengue incidence in Yogyakarta,

Key words and phrases. As treated, cluster randomized trial, GEE, GLMM, per-protocol, test-negative design, dengue, Wolbachia.

Indonesia: Study protocol for a cluster randomised controlled trial. *Trials* **19** 302. https://doi.org/10.1186/S13063-018-2670-Z

BEGG, M. D. and PARIDES, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat. Med.* **22** 2591–2602. https://doi.org/10.1002/SIM.1524

CAVANY, S. M., HUBER, J. H., WIELER, A., ELLIOTT, M., TRAN, Q. M., ESPAÑA, G., MOORE, S. M. and PERKINS, T. A. (2021). Ignoring transmission dynamics leads to underestimation of the impact of a novel intervention against mosquito-borne disease. MedRxiv 2021.11.19.21266602. https://doi.org/10.1101/2021.11.19.21266602

DANIEL, R., ZHANG, J. and FAREWELL, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biom. J.* **63** 528–557. MR4226593 https://doi.org/10.1002/bimj.201900297

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

DUFAULT, S. M. and JEWELL, N. P. (2020). Analysis of counts for cluster randomized trials: Negative controls and test-negative designs. *Stat. Med.* **39** 1429–1439. MR4098500 https://doi.org/10.1002/sim.8488

DUFAULT, S. M., TANAMAS, S. K., INDRIANI, C., AHMAD, C. R. A., UTARINI, A., JEWELL, N. P., SIMMONS, C. P. and ANDERS, K. L. (2023). Reanalysis of cluster randomized trial data to account for exposure misclassification using a per-protocol and as-treated approach. Submitted for Publication.

FOGELSON, A. I, LANDSIEDEL, K. E, DUFAULT, S. M and JEWELL, N. P (2024). Supplement to "As treated analyses of cluster randomized trials." https://doi.org/10.1214/23-AOAS1846SUPP

JACKSON, M. L. and NELSON, J. C. (2013). The test-negative design for estimating influenza vaccine effectiveness. *Vaccine* **31** 2165–2168. https://doi.org/10.1016/J.VACCINE.2013.02.053

JEWELL, N. P., DUFAULT, S., CUTCHER, Z., SIMMONS, C. P. and ANDERS, K. L. (2019). Analysis of cluster-randomized test-negative designs: Cluster-level methods. *Biostatistics* **20** 332–346. MR3922137 https://doi.org/10.1093/biostatistics/kxy005

JOHNSON, K. N. (2015). The impact of Wolbachia on virus infection in mosquitoes. *Viruses* **7** 5705–5717. https://doi.org/10.3390/V7112903

KANG, H. and KEELE, L. (2018). Estimation methods for cluster randomized trials with noncompliance: A study of a biometric smartcard payment system in India. https://doi.org/10.48550/arXiv.1805.03744

NEUHAUS, J. M. and KALBFLEISCH, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54** 638. https://doi.org/10.2307/3109770

PEPE, M. S. 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. https://doi.org/10.1080/03610919408813210

SHRIER, I., STEELE, R. J., VERHAGEN, E., HERBERT, R., RIDDELL, C. A. and KAUFMAN, J. S. (2014). Beyond intention to treat: What is the right question? *Clin. Trials* **11** 28–37. https://doi.org/10.1177/1740774513504151

SMITH, V. A., COFFMAN, C. J. and HUDGENS, M. G. (2021). Interpreting the results of intention-to-treat, per-protocol, and as-treated analyses of clinical trials. *JAMA* **326** 433–434. https://doi.org/10.1001/JAMA.2021.2825

STRAM, D. and LEE, J. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

SULLIVAN, S. G., TCHETGEN, E. J. T. and COWLING, B. J. (2016). Theoretical basis of the test-negative study design for assessment of influenza vaccine effectiveness. *Amer. J. Epidemiol.* **184** 345–353. https://doi.org/10.1093/AJE/KWW064

UTARINI, A., INDRIANI, C., AHMAD, R. A., TANTOWIJOYO, W., ARGUNI, E., ANSARI, M. R., SUPRIYATI, E., WARDANA, D. S., MEITIKA, Y. et al. (2021). Efficacy of Wolbachia-infected mosquito deployments for the control of Dengue. *N. Engl. J. Med.* **384** 2177–2186. https://doi.org/10.1056/NEJMOA2030243

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1519–1542
https://doi.org/10.1214/23-AOAS1847

MODELING EXTREMAL STREAMFLOW USING DEEP LEARNING APPROXIMATIONS AND A FLEXIBLE SPATIAL PROCESS

BY REETAM MAJUMDER[1,a], BRIAN J. REICH[2,b] AND BENJAMIN A. SHABY[3,c]

[1]*Southeast Climate Adaptation Science Center, North Carolina State University,* [a]*rmajumd3@ncsu.edu*
[2]*Department of Statistics, North Carolina State University,* [b]*bjreich@ncsu.edu*
[3]*Department of Statistics, Colorado State University,* [c]*bshaby@colostate.edu*

Quantifying changes in the probability and magnitude of extreme flooding events is key to mitigating their impacts. While hydrodynamic data are inherently spatially dependent, traditional spatial models, such as Gaussian processes, are poorly suited for modeling extreme events. Spatial extreme value models with more realistic tail dependence characteristics are under active development. They are theoretically justified but give intractable likelihoods, making computation challenging for small datasets and prohibitive for continental-scale studies. We propose a process mixture model (PMM) which specifies spatial dependence in extreme values as a convex combination of a Gaussian process and a max-stable process, yielding desirable tail dependence properties but intractable likelihoods. To address this, we employ a unique computational strategy where a feed-forward neural network is embedded in a density regression model to approximate the conditional distribution at one spatial location, given a set of neighbors. We then use this univariate density function to approximate the joint likelihood for all locations by way of a Vecchia approximation. The PMM is used to analyze changes in annual maximum streamflow within the U.S. over the last 50 years and is able to detect areas which show increases in extreme streamflow over time.

REFERENCES

ABRAHAMOWICZ, M., CLAMPL, A. and RAMSAY, J. O. (1992). Nonparametric density estimation for censored survival data: Regression-spline approach. *Canad. J. Statist.* **20** 171–185.

ARCHFIELD, S. A., HIRSCH, R. M., VIGLIONE, A. and BLÖSCHL, G. (2016). Fragmented patterns of flood change across the United States. *Geophys. Res. Lett.* **43** 10–232.

ASADI, P., DAVISON, A. C. and ENGELKE, S. (2015). Extremes on river networks. *Ann. Appl. Stat.* **9** 2023–2050.

BLÖSCHL, G., HALL, J., VIGLIONE, A., PERDIGÃO, R. A., PARAJKA, J., MERZ, B., LUN, D., ARHEIMER, B., ARONICA, G. T. et al. (2019). Changing climate both increases and decreases European river floods. *Nature* **573** 108–111.

BOPP, G. P., SHABY, B. A. and HUSER, R. (2021). A hierarchical max-infinitely divisible spatial model for extreme precipitation. *J. Amer. Statist. Assoc.* **116** 93–106.

BROWN, B. M. and RESNICK, S. I. (1977). Extreme values of independent stochastic processes. *J. Appl. Probab.* **14** 732–739.

CASTRUCCIO, S., HUSER, R. and GENTON, M. G. (2016). High-order composite likelihood inference for max-stable distributions and processes. *J. Comput. Graph. Statist.* **25** 1212–1229.

CHUI, C., SMITH, P. and WARD, J. (1980). Degree of L_p approximation by monotone splines. *SIAM J. Math. Anal.* **11** 436–447.

COLES, S., BAWA, J., TRENNER, L. and DORAZIO, P. (2001). *An Introduction to Statistical Modeling of Extreme Values* **208**. Springer.

CONDON, L., GANGOPADHYAY, S. and PRUITT, T. (2015). Climate change and non-stationary flood risk for the upper Truckee River basin. *Hydrol. Earth Syst. Sci.* **19** 159–175.

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

DAWDY, D. R., GRIFFIS, V. W. and GUPTA, V. K. (2012). Regional flood-frequency analysis: How we got here and where we are going. *J. Hydrol. Eng.* **17** 953–959.

Key words and phrases. Gaussian process, max-stable process, neural networks, spatial extremes, Vecchia approximation.

DE CICCO, L. A., LORENZ, D., HIRSCH, R. M., WATKINS, W. and JOHNSON, M. (2022). dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services, Reston, VA. https://doi.org/10.5066/P9X4L3GE

DE HAAN, L. (1984). A spectral representation for max-stable processes. *Ann. Probab.* **12** 1194–1204.

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction*. Springer, Berlin.

ENGELKE, S. and HITZ, A. S. (2020). Graphical models for extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 871–932. https://doi.org/10.1111/rssb.12355

ERHARDT, R. J. and SMITH, R. L. (2012). Approximate Bayesian computing for spatial extremes. *Comput. Statist. Data Anal.* **56** 1468–1481.

FRANÇOIS, B., SCHLEF, K., WI, S. and BROWN, C. (2019). Design considerations for riverine floods in a changing climate—A review. *J. Hydrol.* **574** 557–573.

FRANKS, S. W. (2002). Identification of a change in climate state using regional flood data. *Hydrol. Earth Syst. Sci.* **6** 11–16.

GERBER, F. and NYCHKA, D. (2021). Fast covariance parameter estimation of spatial Gaussian process models using neural networks. *Stat* **10** e382.

GREENBERG, D., NONNENMACHER, M. and MACKE, J. (2019). Automatic posterior transformation for likelihood-free inference. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 2404–2414.

GUINNESS, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics* **60** 415–429. https://doi.org/10.1080/00401706.2018.1437476

HAZRA, A., HUSER, R. and BOLIN, D. (2021). Realistic and fast modeling of spatial extremes over large geographical domains. https://doi.org/10.48550/ARXIV.2112.10248

HEFFERNAN, J. E. and TAWN, J. A. (2001). Extreme value analysis of a large designed experiment: A case study in bulk carrier safety. *Extremes* **4** 359–378.

HIRABAYASHI, Y., MAHENDRAN, R., KOIRALA, S., KONOSHIMA, L., YAMAZAKI, D., WATANABE, S., KIM, H. and KANAE, S. (2013). Global flood risk under climate change. *Nat. Clim. Change* **3** 816–821.

HIRSCH, R. M. (2011). A perspective on nonstationarity and water management 1. *J. Am. Water Resour. Assoc.* **47** 436–446.

HIRSCH, R. M. and RYBERG, K. R. (2012). Has the magnitude of floods across the USA changed with global CO2 levels? *Hydrol. Sci. J.* **57** 1–9.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** 359–366.

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

HUSER, R., DAVISON, A. C. and GENTON, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes* **19** 79–103.

HUSER, R., DOMBRY, C., RIBATET, M. and GENTON, M. G. (2019). Full likelihood inference for max-stable data. *Stat* **8** e218.

HUSER, R., STEIN, M. L. and ZHONG, P. (2022). Vecchia likelihood approximation for accurate and fast inference in intractable spatial extremes models. Preprint. Available at arXiv:2203.05626.

HUSER, R. and WADSWORTH, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *J. Amer. Statist. Assoc.* **114** 434–444.

JÄRVENPÄÄ, M., GUTMANN, M. U., VEHTARI, A. and MARTTINEN, P. (2021). Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian Anal.* **16** 147–178.

JOE, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*, 1st ed. CRC Press/CRC, Boca Raton.

KABLUCHKO, Z., SCHLATHER, M. and DE HAAN, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Probab.* **37** 2042–2065.

KATZFUSS, M. and GUINNESS, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statist. Sci.* **36** 124–141.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KNOX, J. C. (1993). Large increases in flood magnitude in response to modest changes in climate. *Nature* **361** 430–432.

KOBYZEV, I., PRINCE, S. J. D. and BRUBAKER, M. A. (2021). Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 3964–3979. https://doi.org/10.1109/TPAMI.2020.2992934

KUNDZEWICZ, Z. W., KANAE, S., SENEVIRATNE, S. I., HANDMER, J., NICHOLLS, N., PEDUZZI, P., MECHLER, R., BOUWER, L. M., ARNELL, N. et al. (2014). Flood risk and climate change: Global and regional perspectives. *Hydrol. Sci. J.* **59** 1–28.

KUNDZEWICZ, Z. W., KRYSANOVA, V., DANKERS, R., HIRABAYASHI, Y., KANAE, S., HATTERMANN, F. F., HUANG, S., MILLY, P. C., STOFFEL, M. et al. (2017). Differences in flood hazard projections in Europe–their causes and consequences for decision making. *Hydrol. Sci. J.* **62** 1–14.

KUNKEL, K. E., KARL, T. R., SQUIRES, M. F., YIN, X., STEGALL, S. T. and EASTERLING, D. R. (2020). Precipitation extremes: Trends and relationships with average precipitation and precipitable water in the contiguous United States. *J. Appl. Meteorol. Climatol.* **59** 125–142.

LEDFORD, A. W. and TAWN, J. A. (1996). Satistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187.

LEDFORD, A. W. and TAWN, J. A. (1997). Modelling dependence within joint tail regions. *J. Roy. Statist. Soc. Ser. B* **59** 475–499.

LENZI, A., BESSAC, J., RUDI, J. and STEIN, M. L. (2021). Neural networks for parameter estimation in intractable models. Preprint. Available at arXiv:2107.14346.

LI, L., HOLBROOK, A., SHAHBABA, B. and BALDI, P. (2019). Neural network gradient Hamiltonian Monte Carlo. *Comput. Statist.* **34** 281–299. https://doi.org/10.1007/s00180-018-00861-z

LIMA, C. H., LALL, U., TROY, T. and DEVINENI, N. (2016). A hierarchical Bayesian GEV model for improving local and regional flood quantile estimates. *J. Hydrol.* **541** 816–823.

LINS, H. F. (2012). USGS hydro-climatic data network 2009 (HCDN-2009). *US Geological Survey Fact Sheet* **3047**.

MAJUMDER, R., REICH, B. J and SHABY, B. A (2024). Supplement to "Modeling extremal streamflow using deep learning approximations and a flexible spatial process." https://doi.org/10.1214/23-AOAS1847SUPPA, https://doi.org/10.1214/23-AOAS1847SUPPB

MEEHL, G. A., ZWIERS, F., EVANS, J., KNUTSON, T., MEARNS, L. and WHETTON, P. (2000). Trends in extreme weather and climate events: Issues related to modeling extremes in projections of future climate change. *Bull. Am. Meteorol. Soc.* **81** 427–436.

MERZ, B., AERTS, J., ARNBJERG-NIELSEN, K., BALDI, M., BECKER, A., BICHET, A., BLÖSCHL, G., BOUWER, L. M., BRAUER, A. et al. (2014). Floods and climate: Emerging perspectives for flood risk assessment and management. *Nat. Hazards Earth Syst. Sci.* **14** 1921–1942.

MILLY, P., BETANCOURT, J., FALKENMARK, M., HIRSCH, R. M., KUNDZEWICZ, Z. W., LETTENMAIER, D. P. and STOUFFER, R. J. (2008). Stationarity is dead: Whither water management? *Earth* **4**.

MILLY, P. C., BETANCOURT, J., FALKENMARK, M., HIRSCH, R. M., KUNDZEWICZ, Z. W., LETTENMAIER, D. P., STOUFFER, R. J., DETTINGER, M. D. and KRYSANOVA, V. (2015). On critiques of "Stationarity is dead: Whither water management?". *Water Resour. Res.* **51** 7785–7789.

MILLY, P. C., DUNNE, K. A. and VECCHIA, A. V. (2005). Global pattern of trends in streamflow and water availability in a changing climate. *Nature* **438** 347–350.

MORRIS, S. A., REICH, B. J. and THIBAUD, E. (2019). Exploration and inference in spatial extremes using empirical basis functions. *J. Agric. Biol. Environ. Stat.* **24** 555–572.

NAIR, V. and HINTON, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. *ICML'*10 807–814. Omnipress, Madison, WI, USA.

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010b). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277.

PAPAMAKARIOS, G., NALISNICK, E., REZENDE, D. J., MOHAMED, S. and LAKSHMINARAYANAN, B. (2021). Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.* **22** 1–64.

PENROSE, M. D. (1992). Semi-min-stable processes. *Ann. Probab.* **20** 1450–1463. https://doi.org/10.1214/aop/1176989700

PRICE, L. F., DROVANDI, C. C., LEE, A. and NOTT, D. J. (2018). Bayesian synthetic likelihood. *J. Comput. Graph. Statist.* **27** 1–11.

RASMUSSEN, C. E. (2003). Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. In *Seventh Valencia International Meeting*, *Dedicated to Dennis V. Lindley* 651–659. Oxford Univ. Press, London.

REICH, B. J. and SHABY, B. A. (2012b). A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6** 1430–1451. https://doi.org/10.1214/12-AOAS591

REICH, B. J., SHABY, B. A. and COOLEY, D. (2013). A hierarchical model for serially-dependent extremes: A study of heat waves in the western US. *J. Agric. Biol. Environ. Stat.* **19** 119–135.

RIBATET, M., COOLEY, D. and DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statist. Sinica* **22** 813–845.

SAINSBURY-DALE, M., ZAMMIT-MANGION, A. and HUSER, R. (2023). Fast optimal estimation with intractable models using permutation-invariant neural networks.

SALAS, J. D. and OBEYSEKERA, J. (2014). Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events. *J. Hydrol. Eng.* **19** 554–568.

SANG, H. and GENTON, M. G. (2014b). Tapered composite likelihood for spatial max-stable models. *Spat. Stat.* **8** 86–103.

SANTOS-FERNANDEZ, E., VER HOEF, J. M., PETERSON, E. E., MCGREE, J., ISAAK, D. J. and MENGERSEN, K. (2022b). Bayesian spatio-temporal models for stream networks. *Comput. Statist. Data Anal.* **170** 107446.

SCHLATHER, M. (2002). Models for stationary max-stable random fields. *Extremes* **5** 33–44.

SHARMA, A., WASKO, C. and LETTENMAIER, D. P. (2018). If precipitation extremes are increasing, why aren't floods? *Water Resour. Res.* **54** 8545–8551.

SMITH, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

SRAJ, M., VIGLIONE, A., PARAJKA, J. and BLÖSCHL, G. (2016). The influence of non-stationarity in extreme hydrological events on flood frequency estimation. *J. Hydrol. Hydromech.* **64** 426–437.

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004b). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 275–296.

TAWN, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika* **77** 245–253.

VECCHIA, A. V. (1988b). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312.

VEHTARI, A., GELMAN, A. and GABRY, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

VOGEL, R. M., YAINDL, C. and WALTER, M. (2011). Nonstationarity: Flood magnification and recurrence reduction factors in the United States. *J. Am. Water Resour. Assoc.* **47** 464–474.

WADSWORTH, J. L. (2015). On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika* **102** 705–711. MRMR3394285 https://doi.org/10.1093/biomet/asv029

WADSWORTH, J. L. and TAWN, J. A. (2012b). Dependence modelling for spatial extremes. *Biometrika* **99** 253–272.

WADSWORTH, J. L. and TAWN, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101** 1–15. MRMR3180654 https://doi.org/10.1093/biomet/ast042

WALTER, M. (2010). Increasing trends in peak flows in the northeastern United States and their impacts on design Ph.D. thesis Tufts Univ.

WANG, H. and LI, J. (2018b). Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions. *Neural Comput.* **30** 3072–3094.

WANG, Y. and STOEV, S. A. (2010). On the structure and representations of max-stable processes. *Adv. in Appl. Probab.* **42** 855–877. https://doi.org/10.1239/aap/1282924066

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

WILKINSON, R. (2014). Accelerating ABC methods using Gaussian processes. In *Artificial Intelligence and Statistics* 1015–1023. PMLR.

WINSEMIUS, H. C., JONGMAN, B., VELDKAMP, T. I., HALLEGATTE, S., BANGALORE, M. and WARD, P. J. (2018). Disaster risk, climate change, and poverty: Assessing the global exposure of poor people to floods and droughts. *Environ. Dev. Econ.* **23** 328–348.

XU, S. and MAJUMDER, R. (2022). SPQR: Semi-Parametric Quantile Regression. R package version 0.1.0.

XU, S. G. and REICH, B. J. (2021). Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics* **00** 1–14. https://doi.org/10.1111/biom.13576

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1543–1564
https://doi.org/10.1214/23-AOAS1848

ASSESSING SCREENING EFFICACY IN THE PRESENCE OF CANCER OVERDIAGNOSIS

BY YING HUANG[a] AND ZIDING FENG[b]

Fred Hutchinson Cancer Center, [a]*yhuang@fredhutch.org,* [b]*zfeng@fredhutch.org*

Cancer screening facilitates the early detection of cancer at a stage when treatment is often most effective. However, it also brings the risk of overdiagnosis, where a diagnosis made through screening would not have led to symptoms or death during the patient's lifetime. In this paper we tackle a significant unresolved issue in the evaluation of screening efficacy: selecting primary endpoints and inferential procedures that efficiently consider potential overdiagnosis in screening trials. This is motivated by the necessity to design and analyze a phase IV Early Detection Initiative (EDI) trial for evaluating a pancreatic cancer screening strategy. We introduce two novel approaches for assessing screening efficacy, grounded on cancer stage shift. These methods address potential overdiagnosis by: (i) borrowing information about clinical diagnosis from the control arm that hasn't undergone screening (the BR approach) and (ii) performing sensitivity analysis, contingent upon a conservative bound of the overdiagnosis magnitude (the SEN-T approach). Analytical methods and extensive simulation studies underscore the superiority of our proposed methods, demonstrating enhanced efficiency in estimating and testing screening efficacy compared to existing methods. The latter either overlook overdiagnosis or adhere to a valid, yet conservative, cumulative incidence endpoint. We illustrate the practical application of these approaches using ovarian cancer data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. The results affirm that our methods bolster an efficient and robust study design for cancer screening trials.

REFERENCES

ANDRIOLE, G. L., CRAWFORD, E. D., GRUBB III, R. L., BUYS, S. S., CHIA, D., CHURCH, T. R., FOUAD, M. N., GELMANN, E. P., KVALE, P. A. et al. (2009). Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.* **360** 1310–1319.

ANDRIOLE, G. L., CRAWFORD, E. D., GRUBB III, R. L., BUYS, S. S., CHIA, D., CHURCH, T. R., FOUAD, M. N., ISAACS, C., KVALE, P. A. et al. (2012). Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: Mortality results after 13 years of follow-up. *J. Natl. Cancer Inst.* **104** 125–132.

BUYS, S. S., PARTRIDGE, E., BLACK, A., JOHNSON, C. C., LAMERATO, L., ISAACS, C., READING, D. J., GREENLEE, R. T., YOKOCHI, L. A. et al. (2011). Effect of screening on ovarian cancer mortality: The Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening randomized controlled trial. *JAMA* **305** 2295–2303.

CARTER, J. L., COLETTI, R. J. and HARRIS, R. P. (2015). Quantifying and monitoring overdiagnosis in cancer screening: A systematic review of methods. *BMJ* **350** g7773. https://doi.org/10.1136/bmj.g7773

CHARI, S. T., MAITRA, A., MATRISIAN, L. M., SHRADER, E. E., WU, B. U., KAMBADAKONE, A., ZHAO, Y.-Q., KENNER, B., RINAUDO, J. A. S. et al. (2021). Early detection initiative: A randomized controlled trial of algorithm-based screening in patients with new onset hyperglycemia and diabetes for early detection of pancreatic ductal adenocarcinoma. *Contemp. Clin. Trials* 106659.

ESSERMAN, L. J., THOMPSON, I. M., REID, B., NELSON, P., RANSOHOFF, D. F., WELCH, H. G., HWANG, S., BERRY, D. A., KINZLER, K. W. et al. (2014). Addressing overdiagnosis and overtreatment in cancer: A prescription for change. *Lancet Oncol.* **15** e234–e242.

Key words and phrases. Biomarker, cancer screening, cumulative incidence, overdiagnosis, sensitivity analysis, stage shift.

ETZIONI, R., GULATI, R., MALLINGER, L. and MANDELBLATT, J. (2013). Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann. Intern. Med.* **158** 831–838. https://doi.org/10.7326/0003-4819-158-11-201306040-00008

HUANG, Y. and FENG, Z. (2024). Supplement to "Assessing screening efficacy in the presence of cancer overdiagnosis." https://doi.org/10.1214/23-AOAS1848SUPP

MOLENBERGHS, G., KENWARD, M. G. and GOETGHEBEUR, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 15–29.

OKEN, M. M., HOCKING, W. G., KVALE, P. A., ANDRIOLE, G. L., BUYS, S. S., CHURCH, T. R., CRAWFORD, E. D., FOUAD, M. N., ISAACS, C. et al. (2011). Screening by chest radiograph and lung cancer mortality: The prostate, lung, colorectal, and ovarian (PLCO) randomized trial. *JAMA* **306** 1865–1873.

PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THORNQUIST, M., WINGET, M. and YASUI, Y. (2001). Phases of biomarker development for early detection of cancer. *J. Natl. Cancer Inst.* **93** 1054–1061. https://doi.org/10.1093/jnci/93.14.1054

PROROK, P. C., ANDRIOLE, G. L., BRESALIER, R. S., BUYS, S. S., CHIA, D., CRAWFORD, E. D., FOGEL, R., GELMANN, E. P., GILBERT, F. et al. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control. Clin. Trials* **21** 273S–309S.

RIPPING, T. M., HAAF, K. T., VERBEEK, A. L. M., VAN RAVESTEYN, N. T. and BROEDERS, M. J. M. (2017). Quantifying overdiagnosis in cancer screening: A systematic review to evaluate the methodology. *J. Natl. Cancer Inst.* **109**. https://doi.org/10.1093/jnci/djx060

SCHOEN, R. E., PINSKY, P. F., WEISSFELD, J. L., YOKOCHI, L. A., CHURCH, T., LAIYEMO, A. O., BRESALIER, R., ANDRIOLE, G. L., BUYS, S. S. et al. (2012). Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *N. Engl. J. Med.* **366** 2345–2357.

SHARMA, A., KANDLAKUNTA, H., NAGPAL, S. J. S., FENG, Z., HOOS, W., PETERSEN, G. M. and CHARI, S. T. (2018). Model to determine risk of pancreatic cancer in patients with new-onset diabetes. *Gastroenterology* **155** 730–739. PMCID: PMC6120785.

SRIVASTAVA, S., KOAY, E. J., BOROWSKY, A. D., DE MARZO, A. M., GHOSH, S., WAGNER, P. D. and KRAMER, B. S. (2019). Cancer overdiagnosis: A biological challenge and clinical dilemma. *Nat. Rev. Cancer* **19** 349–358.

TEAM, N. L. S. T. R. (2011). The national lung screening trial: Overview and study design. *Radiology* **258** 243–253.

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311

WELCH, H. G. and BLACK, W. C. (2010). Overdiagnosis in cancer. *J. Natl. Cancer Inst.* **102** 605–613.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1565–1595
https://doi.org/10.1214/23-AOAS1849

A BAYESIAN HIERARCHICAL SMALL AREA POPULATION MODEL ACCOUNTING FOR DATA SOURCE SPECIFIC METHODOLOGIES FROM AMERICAN COMMUNITY SURVEY, POPULATION ESTIMATES PROGRAM, AND DECENNIAL CENSUS DATA

BY EMILY N. PETERSON[1,a], RACHEL C. NETHERY[2,c], TULLIA PADELLINI[3,f], JARVIS T. CHEN[2,d], BRENT A. COULL[2,e], FRÉDÉRIC B. PIEL[3,g], JON WAKEFIELD[4,i], MARTA BLANGIARDO[3,h] AND LANCE A. WALLER[1,b]

[1]*Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University,* [a]*emily.nancy.peterson@emory.edu,* [b]*lwaller@emory.edu*

[2]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [c]*rnethery@hsph.harvard.edu,* [d]*jarvis@hsph.harvard.edu,* [e]*jarvis@hsph.harvard.edu*

[3]*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London,* [f]*tullia.padellini@bancaditalia.it,* [g]*jarvis@hsph.harvard.edu,* [h]*m.blangiardo@imperial.ac.uk*

[4]*Department of Biostatistics, School of Public Health, University of Washington,* [i]*jonno@uw.edu*

Small area population counts are necessary for many epidemiological studies, yet their quality and accuracy are often not assessed. In the United States, small area population counts are published by the United States Census Bureau (USCB) in the form of the decennial census counts, intercensal population projections (PEP), and American Community Survey (ACS) estimates. Although there are significant relationships between these three data sources, there are important contrasts in data collection, data availability, and processing methodologies such that each set of reported population counts may be subject to different sources and magnitudes of error. Additionally, these data sources do not report identical small area population counts due to post-survey adjustments specific to each data source. Consequently, in public health studies, small area disease/mortality rates may differ depending on which data source is used for denominator data. To accurately estimate annual small area population counts *and their* associated uncertainties, we present a Bayesian population (BPop) model, which fuses information from all three USCB sources, accounting for data source specific methodologies and associated errors. We produce comprehensive small area race-stratified estimates of the true population, and associated uncertainties, given the observed trends in all three USCB population estimates. The main features of our framework are: (1) a single model integrating multiple data sources, (2) accounting for data source specific data generating mechanisms and specifically accounting for data source specific errors, and (3) prediction of population counts for years without USCB reported data. We focus our study on the Black and White only populations for 159 counties of Georgia and produce estimates for years 2006–2023. We compare BPop population estimates to decennial census counts, PEP annual counts, and ACS multi-year estimates. Additionally, we illustrate and explain the different types of data source specific errors. Lastly, we compare model performance using simulations and validation exercises. Our Bayesian population model can be extended to other applications at smaller spatial granularity and for demographic subpopulations defined further by race, age, and sex, and/or for other geographical regions.

REFERENCES

BLANGIARDO, M. and CAMELETTI, M. (2015). *Spatial and spatiotemporal Bayesian Models with R-INLA*. Wiley, Chichester. MR3364017

Key words and phrases. Small area population estimates, American Community Survey, decennial census, sampling and nonsampling errors, Bayesian methods.

BRADLEY, J., WIKLE, C. and HOLAN, S. (2015). Spatiotemporal change of support with application to american community survey multi-year period estimates. Available at https://arxiv.org/pdf/1508.01451.pdf.

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2015). Multivariate spatiotemporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Ann. Appl. Stat.* **9** 1761–1791. MR3456353 https://doi.org/10.1214/15-AOAS862

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2016). Multivariate spatiotemporal survey fusion with application to the American Community Survey and Local Area Unemployment Statistics. *Stat* **5** 224–233. MR3564657 https://doi.org/10.1002/sta4.120

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2018). Computationally efficient multivariate spatiotemporal models for high-dimensional count-valued data (with discussion). *Bayesian Anal.* **13** 253–310. MR3773410 https://doi.org/10.1214/17-BA1069

CAMELETTI, M., LINDGREN, F., SIMPSON, D. and RUE, H. (2013). Spatiotemporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv. Stat. Anal.* **97** 109–131. MR3045763 https://doi.org/10.1007/s10182-012-0196-3

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatiotemporal Data*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2848400

DAVIS, R. A., FOKIANOS, K., HOLAN, S. H., JOE, H., LIVSEY, J., LUND, R., PIPIRAS, V. and RAVISHANKER, N. (2021). Count time series: A methodological review. *J. Amer. Statist. Assoc.* **116** 1533–1547. MR4309291 https://doi.org/10.1080/01621459.2021.1904957

DE VALPINE, P., TUREK, D., PACIOREK, C. J., ANDERSON-BERGMAN, C., TEMPLE LANG, D. and BODIK, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *J. Comput. Graph. Statist.* **26** 403–413. MR3640196 https://doi.org/10.1080/10618600.2016.1172487

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

KNORR-HELD, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Stat. Med.* **19** 2555–2567.

LAWSON, A. B. (2013). *Bayesian Disease Mapping*: *Hierarchical Modeling in Spatial Epidemiology*. *Interdisciplinary Statistics*. CRC Press, Boca Raton, FL. MR2484272

NATIONAL ADVISORY COMMITTEE – 2022 SPRING VIRTUAL MEETING (2022). Blended base for population estimates. Available at https://www2.census.gov/about/partners/cac/nac/meetings/2022-05/presentation-blended-base-for-population-estimates.pdf. Accessed: 07-01-2022.

NETHERY, R., RUSHOVICH, T., PETERSON, E., CHEN, J., WATERMAN, P., KRIEGER, N., WALLER, L. and COULL, B. (2021). Comparing the performance of three census tract denominator sources for real-time disease incidence modeling: US Decennial census, American Community Survey, and Worldpop. *Health Place*.

PETERSON, E. N., NETHERY, R. C., PADELLINI, T., CHEN, J. T., COULL, B. A., PIEL, F. B., WAKEFIELD, J., BLANGIARDO, M. and WALLER, L. A. (2024). Supplement to "A Bayesian hierarchical small area population model accounting for data source specific methodologies from American Community Survey, Population Estimates Program, and Decennial census data." https://doi.org/10.1214/23-AOAS1849SUPP

PLUMMER, M. (2017). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

POPULATION ESTIMATION PROGRAM, U.S. CENSUS BUREAU (2019). Methodology for the United States population estimates: Vintage 2019. Available at https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2019/natstcopr-methv2.pdf. Accessed: 07-05-2021.

POPULATION ESTIMATION PROGRAM, U.S. CENSUS BUREAU (2022). Methodology for the United States population estimates: Vintage 2022. Available at https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-detail.html. Accessed: 07-20-2022.

PRESTON, S., HEUVELINE, P. and GUILLOT, M. (2000). *Demography*; *Measuring and Modeling Population Processes*, 1st ed. Wiley Blackwell, New York.

RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Amer. Statist. Assoc.* **102** 474–486. MR2370848 https://doi.org/10.1198/016214506000001293

RIEBLER, A., SØRBYE, S. H., SIMPSON, D. and RUE, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25** 1145–1165. MR3541089 https://doi.org/10.1177/0962280216660421

SPIELMAN, S. and FOLCH, D. (2015). Reducing uncertainty in the American Community Survey through data-driven regionalization. *PLoS ONE* **10** 1–21.
SPIELMAN, S. E., FOLCH, D. and NAGLE, N. (2014). Patterns and causes of uncertainty in the American Community Survey. *Appl. Geogr.* **46** 147–157. https://doi.org/10.1016/j.apgeog.2013.11.002
STARSINIC, M. and TERSINE, A. JR. (2007). Analysis of variance estimates from American Community Survey multi-year estimates. In *Proceedings of the Section of Survey Research Methods* 3011–3017. American Statistical Association, Alexandria, VA.
SU, Y.-S. and YAJIMA, M. (2020). R2jags: Using R to run 'JAGS'. Available at https://CRAN.R-project.org/package=R2jags. Accessed: 2020-07-09.
UNITED STATES CENSUS BUREAU (2012). Decennial census: Complete technical documentation.
U.S. CENSUS BUREAU (2004). Accuracy and coverage evaluation of census 2000. Available at https://www2.census.gov/programs-surveys/decennial/2000/technical-documentation/coverage-evaluation/dssd03-dm.pdf.
U.S. CENSUS BUREAU (2009). A compass for understanding and using American Community Survey data: What researchers need to know. Available at https://www.census.gov/content/dam/Census/library/publications.2009/acs/ACSResearch.pdf.
U.S. CENSUS BUREAU (2014a). American Community Survey: Design and methodology Chapter 11: Weighting and estimation. Available at https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2022/acs_design_methodology_ch11_2022.pdf.
U.S. CENSUS BUREAU (2014b). The American Community Survey: Design and methodology Chapter 4: Sample design and selection. Available at https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2014/acs_design_methodology_ch04_2014.pdf.
U.S. CENSUS BUREAU (2018). Understanding and using American Community Survey data: What all data users need to know. Available at https://www.census.gov/programs-surveys/acs/library/handbooks/general.html.
U.S. CENSUS BUREAU (2020). 2020: Dec decennial post-enumeration survey. Available at https://data.census.gov/cedsci/table?y=2020&d=DEC%20Decennial%20Post-Enumeration%20Survey&tid=DECENNIALPES2020.C_RACEHISUS. Accessed: 08-10-2023.
U.S. CENSUS BUREAU (2022). Census coverage estimates for people in the United States by state and census operations: 2020 post enumeration survey estimation report. Available at https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/census-coverage-estimates-for-people-in-the-united-states-by-state-and-census-operations.pdf. Accessed: 08-10-2023.
U.S. CENSUS BUREAU: CCM (2012). Estimates of undercount and overcount in the 2010 census. Available at https://www.census.gov/newsroom/releases/archives/2010census/cb12-95.html. Accessed: 05-15-2021.
U.S. CENSUS BUREAU: MEASURES OF NONSAMPLING ERROR (2015). Statistical quality standard d3: Producing measures and indicators of nonsampling error. Available at https://www.census.gov/about/policies/quality/standards/standardd3.html.
U.S. DEPARTMENT OF COMMERCE (2012). Dssd 2010 census coverage measurement memorandum series #2010-g-01. Available at https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/methodology/g-series/g01.pdf. Accessed: 08-10-2023.
WAKEFIELD, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8** 158–183.
WALKER, K. (2020). tidycensus: Load US census boundary and attribute data as 'tidyverse'. r package version 0.9.9.2. Available at https://walker-data.com/tidycensus/articles/basic-usage.html. Accessed: 08-10-2022.
WALLER, L. A., CARLIN, B. P., XIA, H. and GELFAND, A. E. (1997). Hierarchical spatiotemporal mapping of diease rates. *J. Amer. Statist. Assoc.* **92** 607–617.
WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data. Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2075123 https://doi.org/10.1002/0471662682
WANG, J. C., HOLAN, S. H., NANDRAM, B., BARBOZA, W., TOTO, C. and ANDERSON, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *J. Agric. Biol. Environ. Stat.* **17** 84–106. MR2912556 https://doi.org/10.1007/s13253-011-0067-5
WORLDPOP (2020). Worldpop griddded population estimate datasets and tools. How are they different and which should I use? Available at https://www.worldpop.org/methods/populations.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1596–1617
https://doi.org/10.1214/23-AOAS1850

SPATIAL PREDICTIONS ON PHYSICALLY CONSTRAINED DOMAINS: APPLICATIONS TO ARCTIC SEA SALINITY DATA

BY BORA JIN[1,a], AMY H. HERRING[2,b] AND DAVID DUNSON[2,c]

[1]*Department of Biostatistics, Johns Hopkins University,* [a]*bjin9@jh.edu*
[2]*Department of Statistical Science, Duke University,* [b]*amy.herring@duke.edu,* [c]*dunson@duke.edu*

In this paper we predict sea surface salinity (SSS) in the Arctic Ocean based on satellite measurements. SSS is a crucial indicator for ongoing changes in the Arctic Ocean and can offer important insights about climate change. We particularly focus on areas of water mistakenly flagged as ice by satellite algorithms. To remove bias in the retrieval of salinity near sea ice, the algorithms use conservative ice masks, which result in considerable loss of data. We aim to produce realistic SSS values for such regions to obtain more complete understanding about the SSS surface over the Arctic Ocean and benefit future applications that may require SSS measurements near edges of sea ice or coasts. We propose a class of scalable nonstationary processes that can handle large data from satellite products and complex geometries of the Arctic Ocean. Barrier overlap-removal acyclic directed graph GP (BORA-GP) constructs sparse directed acyclic graphs (DAGs) with neighbors conforming to barriers and boundaries, enabling characterization of dependence in constrained domains. The BORA-GP models produce more sensible SSS values in regions without satellite measurements and show improved performance in various constrained domains in simulation studies compared to state-of-the-art alternatives. An R package is available at https://github.com/jinbora0720/boraGP.

REFERENCES

AMAP (1998). AMAP Assessment Report: Arctic pollution issues. Technical Report, Arctic Monitoring and Assessment Programme (AMAP).

BAKKA, H., VANHATALO, J., ILLIAN, J. B., SIMPSON, D. and RUE, H. (2019). Non-stationary Gaussian models with physical barriers. *Spat. Stat.* **29** 268–288. MR3903698 https://doi.org/10.1016/j.spasta.2019.01.002

CARMACK, E., POLYAKOV, I., PADMAN, L., FER, I., HUNKE, E., HUTCHINGS, J., JACKSON, J., KELLEY, D., KWOK, R. et al. (2015). Toward quantifying the increasing role of oceanic heat in sea ice loss in the new Arctic. *Bull. Amer. Meteorol. Soc.* **96** 2079–2105. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society. https://doi.org/10.1175/BAMS-D-13-00177.1

CAVALIERI, D. J. and PARKINSON, C. L. (2012). Arctic sea ice variability and trends, 1979–2010. *Cryosphere*. **6** 881–889. Publisher: Copernicus GmbH. https://doi.org/10.5194/tc-6-881-2012

DAI, N., KANG, H., JONES, G. L. and FIECAS, M. B. (2021). A Bayesian latent spatial model for mapping the cortical signature of progression to Alzheimer's disease. *Canad. J. Statist.* **49** 46–62. MR4238354 https://doi.org/10.1002/cjs.11588

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* **111** 800–812. MR3538706 https://doi.org/10.1080/01621459.2015.1044091

DATTA, A., BANERJEE, S., FINLEY, A. O. and GELFAND, A. E. (2016b). On nearest-neighbor Gaussian process models for massive spatial data. *Wiley Interdiscip. Rev.: Comput. Stat.* **8** 162–171. MR3544254 https://doi.org/10.1002/wics.1383

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016c). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. MR3553225 https://doi.org/10.1214/16-AOAS931

Key words and phrases. Arctic Ocean, barriers, directed acyclic graphs, sea surface salinity, SMAP, spatial statistics.

DAVIS, A. P., GRONDIN, C. J., JOHNSON, R. J., SCIAKY, D., MCMORRAN, R., WIEGERS, J., WIEGERS, T. C. and MATTINGLY, C. J. (2019). The comparative toxicogenomics database: Update 2019. *Nucleic Acids Res.* **47** D948–D954. https://doi.org/10.1093/nar/gky868

DIEBOLD, F. X., GÖBEL, M., GOULET COULOMBE, P., RUDEBUSCH, G. D. and ZHANG, B. (2021). Optimal combination of Arctic sea ice extent measures: a dynamic factor modeling approach. *Int. J. Forecast.* **37** 1509–1519. MR4614577 https://doi.org/10.1016/j.ijforecast.2020.10.006

DUNSON, D. B., WU, H.-T. and WU, N. (2022). Graph based Gaussian processes on restricted domains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 414–439. MR4412992 https://doi.org/10.1111/rssb.12486

EMERY, X. (2009). The Kriging update equations and their application to the selection of neighboring data. *Comput. Geosci.* **13** 269–280. https://doi.org/10.1007/s10596-008-9116-8

FARMER, J. R., SIGMAN, D. M., GRANGER, J., UNDERWOOD, O. M., FRIPIAT, F. and CRONIN, T. M. (2021). Arctic Ocean stratification set by sea level and freshwater inputs since the last ice age. *Nat. Geosci.* **14** 684–689, 9. Publisher: Nature Publishing Group. https://doi.org/10.1038/s41561-021-00789-y

FERAGEN, A., LAUZE, F. and HAUBERG, S. (2015). Geodesic exponential kernels: When curvature and linearity conflict. In 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 3032–3042. IEEE Press, Boston. https://doi.org/10.1109/CVPR.2015.7298922

FETTERER, F., KNOWLES, K., MEIER, W. N., SAVOIE, M. and WINDNAGEL, A. K. (2017). Sea ice index. Version 3. https://doi.org/10.7265/N5K072F8

FINLEY, A. O., DATTA, A., COOK, B. D., MORTON, D. C., ANDERSEN, H. E. and BANERJEE, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Statist.* **28** 401–414. MR3974889 https://doi.org/10.1080/10618600.2018.1537924

FORE, A. G., YUEH, S. H., TANG, W., STILES, B. and HAYASHI, A. K. (2016). Combined active/passive retrievals of ocean vector wind and sea surface salinity with SMAP. *IEEE Trans. Geosci. Remote Sens.* **54** 7396–7404. https://doi.org/10.1109/TGRS.2016.2601486

FOURNIER, S., LEE, T., TANG, W., STEELE, M. and OLMEDO, E. (2019). Evaluation and intercomparison of SMOS, aquarius, and SMAP sea surface salinity products in the Arctic Ocean. *Remote Sens.* **11** 3043, 24. Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs11243043

FOURNIER, S., LEE, T., WANG, X., ARMITAGE, T. W. K., WANG, O., FUKUMORI, I. and KWOK, R. (2020). Sea surface salinity as a proxy for Arctic Ocean freshwater changes. *J. Geophys. Res., Oceans* **125** e2020JC016110. https://doi.org/10.1029/2020JC016110

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395 https://doi.org/10.1080/10618600.2014.914442

GUINOTTE, J. M. and FABRY, V. J. (2008). Ocean acidification and its potential effects on marine ecosystems. *Ann. N.Y. Acad. Sci.* **1134** 320–342. https://doi.org/10.1196/annals.1439.013

HAINE, T. W. N., CURRY, B., GERDES, R., HANSEN, E., KARCHER, M., LEE, C., RUDELS, B., SPREEN, G., DE STEUR, L. et al. (2015). Arctic freshwater export: Status, mechanisms, and prospects. *Glob. Planet. Change* **125** 13–35. https://doi.org/10.1016/j.gloplacha.2014.11.013

HASSOL, S. J. (2004). *Impacts of a Warming Arctic*: *Arctic Climate Impact Assessment*. Cambridge Univ. Press, Cambridge. OCLC: ocm56942125.

HOEGH-GULDBERG, O., MUMBY, P. J., HOOTEN, A. J., STENECK, R. S., GREENFIELD, P., GOMEZ, E., HARVELL, C. D., SALE, P. F., EDWARDS, A. J. et al. (2007). Coral reefs under rapid climate change and ocean acidification. *Science* **318** 1737–1742. Publisher: American Association for the Advancement of Science. https://doi.org/10.1126/science.1152509

JIN, B., HERRING, A. H. and DUNSON, D. (2024). Supplement to "Spatial predictions on physically constrained domains: Applications to Arctic sea salinity data." https://doi.org/10.1214/23-AOAS1850SUPPA, https://doi.org/10.1214/23-AOAS1850SUPPB, https://doi.org/10.1214/23-AOAS1850SUPPC

KERR, Y., RODRIGUEZ-FERNANDEZ, N., ENTEKHABI, D., BINDLISH, R., LEE, T., YUEH, S., LAGERLOEF, G., PIERRE WIGNERON, J., BOUTIN, J., et al. (2018). Present and future of L-band radiometry. In *IGARSS* 2018–2018 *IEEE International Geoscience and Remote Sensing Symposium* 1994–1997. https://doi.org/10.1109/IGARSS.2018.8517457

KIRSANOV, D. (2021). Exact geodesic for triangular meshes. MATLAB central file exchange.

LI, D. and DUNSON, D. B. (2020). Geodesic distance estimation with spherelets. Available at arXiv:1907.00296.

LILA, E., SANGALLI, L. M., ARNONE, E., RAMSAY, J. and FORMAGGIA, L. (2020). fdaPDE: Statistical analysis of functional and spatial data, based on regression with PDE regularization. R package version 1.0-9.

LIN, L. and DUNSON, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349 https://doi.org/10.1093/biomet/ast063

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. With discussion and a reply by the authors. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

LØLAND, A. and HØST, G. (2003). Spatial covariance modelling in a complex coastal domain by multidimensional scaling. *Environmetrics* **14** 307–321. https://doi.org/10.1002/env.588

MÄKINEN, J. and VANHATALO, J. (2016). Hydrographic responses to regional covariates across the Kara Sea. *J. Geophys. Res., Oceans* **121** 8872–8887. https://doi.org/10.1002/2016JC011981

MEISSNER, T. and MANASTER, A. (2021). SMAP salinity retrievals near the sea-ice edge using multi-channel AMSR2 brightness temperatures. *Remote Sens.* **13** 5120, 24. Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs13245120

NEELON, B. and DUNSON, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60** 398–406. MR2066274 https://doi.org/10.1111/j.0006-341X.2004.00184.x

NGHIEM, S. V., HALL, D. K., RIGOR, I. G., LI, P. and NEUMANN, G. (2014). Effects of Mackenzie River discharge and bathymetry on sea ice in the Beaufort Sea. *Geophys. Res. Lett.* **41** 873–879. https://doi.org/10.1002/2013GL058956

NIU, M., CHEUNG, P., LIN, L., DAI, Z., LAWRENCE, N. and DUNSON, D. (2019). Intrinsic Gaussian processes on complex constrained domains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 603–627. MR3961500 https://doi.org/10.1111/rssb.12320

PERUZZI, M. and DUNSON, D. B. (2022). Spatial meshing for general Bayesian multivariate models. Available at arXiv:2201.10080.

POLUKHIN, A. (2019). The role of river runoff in the Kara Sea surface layer acidification and carbonate system changes. *Environ. Res. Lett.* **14** 105007. Publisher: IOP Publishing. https://doi.org/10.1088/1748-9326/ab421e

POLYAKOV, I. V., PNYUSHKOV, A. V. and TIMOKHOV, L. A. (2012). Warming of the intermediate Atlantic water of the Arctic Ocean in the 2000s. *J. Climate* **25** 8362–8370. Publisher: American Meteorological Society Section: Journal of Climate. https://doi.org/0.1175/JCLI-D-12-00266.1

RAMSAY, T. (2002). Spline smoothing over difficult regions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 307–319. MR1904707 https://doi.org/10.1111/1467-9868.00339

RATHBUN, S. L. (1998). Spatial modelling in irregularly shaped regions: Kriging estuaries. *Environmetrics.* **9** 109–129. https://doi.org/10.1002/(SICI)1099-095X(199803/04)9:2<109::AID-ENV279>3.0.CO;2-L

REUL, N., GRODSKY, S. A., ARIAS, M., BOUTIN, J., CATANY, R., CHAPRON, B., D'AMICO, F., DINNAT, E., DONLON, C. et al. (2020). Sea surface salinity estimates from spaceborne L-band radiometers: An overview of the first decade of observation (2010–2019). *Remote Sens. Environ.* **242** 111769. https://doi.org/10.1016/j.rse.2020.111769

REYNOLDS, R. W., ZHANG, H., SMITH, T. M., GENTEMANN, C. L. and WENTZ, F. (2005). Impacts of in situ and additional satellite data on the accuracy of a sea-surface temperature analysis for climate. *Int. J. Climatol.* **25** 857–864. https://doi.org/10.1002/joc.1168

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

SANGALLI, L. M. (2021). Spatial regression with partial differential equation regularisation. *Int. Stat. Rev.* **89** 505–531. MR4411916 https://doi.org/10.1111/insr.12444

SANGALLI, L. M., RAMSAY, J. O. and RAMSAY, T. O. (2013). Spatial spline regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 681–703. MR3091654 https://doi.org/10.1111/rssb.12009

SHESTAKOVA, A. A., TOROPOV, P. A. and MATVEEVA, T. A. (2020). Climatology of extreme downslope windstorms in the Russian Arctic. *Weather Clim. Extrem.* **28** 100256. https://doi.org/10.1016/j.wace.2020.100256

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 275–296. MR2062376 https://doi.org/10.1046/j.1369-7412.2003.05512.x

TANG, W., YUEH, S., YANG, D., FORE, A., HAYASHI, A., LEE, T., FOURNIER, S. and HOLT, B. (2018). The potential and challenges of using soil moisture active passive (SMAP) sea surface salinity to monitor Arctic Ocean freshwater changes. *Remote Sens.* **10** 869, 6. Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs10060869

TANG, W., YUEH, S. H., FORE, A. G., HAYASHI, A. and STEELE, M. (2021). An empirical algorithm for mitigating the sea ice effect in SMAP radiometer for sea surface salinity retrieval in the Arctic seas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14** 11986–11997. https://doi.org/10.1109/JSTARS.2021.3127470

VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. MR0964183 https://doi.org/10.1111/j.2517-6161.1988.tb01729.x

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734 https://doi.org/10.1111/j.1467-9868.2010.00749.x

WOOD, S. N., BRAVINGTON, M. V. and HEDLEY, S. L. (2008). Soap film smoothing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 931–955. MR2530324 https://doi.org/10.1111/j.1467-9868.2008.00665.x

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

ZHANG, J. and LIN, L. (2018). Bounded regression with Gaussian process projection. Available at arXiv:1810.11881.

ZHANG, L. and BANERJEE, S. (2022). Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data. *Biometrics* **78** 560–573. MR4450576 https://doi.org/10.1111/biom.13452

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1618–1641
https://doi.org/10.1214/23-AOAS1851

A FRAMEWORK FOR ANALYSING LONGITUDINAL DATA INVOLVING TIME-VARYING COVARIATES

BY REZA DRIKVANDI[1,a], GEERT VERBEKE[2,b] AND GEERT MOLENBERGHS[3,c]

[1]*Department of Mathematical Sciences, Durham University,* [a]*reza.drikvandi@durham.ac.uk*
[2]*I-BioStat, KU Leuven,* [b]*geert.verbeke@kuleuven.be*
[3]*I-BioStat, Universiteit Hasselt,* [c]*geert.molenberghs@uhasselt.be*

Standard models for longitudinal data ignore the stochastic nature of time-varying covariates and their stochastic evolution over time by treating them as fixed variables. There have been recent methods for modelling time-varying covariates; however, those methods cannot be applied to analyse longitudinal data when the longitudinal response and the time-varying covariates for each subject are measured at different time points. Moreover, it is difficult to study the temporal effects of a time-varying covariate on the longitudinal response and the temporal correlation between them. Motivated by data from an AIDS cohort study conducted over 26 years at the University Hospitals Leuven in which the measurements on the CD4 cell count and viral load for patients are not taken at the same time point, we present a framework to address those challenges by using joint multivariate mixed models to jointly model time-varying covariates and a longitudinal response, instead of including time-varying covariates in the response model. This approach also has the advantage that one can study the association between the covariate at any time point and the response at any other time point without having to explicitly model the conditional distribution of the response given the covariate. We use penalised spline functions of time to capture the evolutions of both the response and time-varying covariates over time.

REFERENCES

BRUMBACK, B. A., RUPPERT, D. and WAND, M. P. (1999). Comment on variable selection and function estimation in additive nonparametric regression using a data-based prior. *J. Amer. Statist. Assoc.* **94** 794–797.

CHEN, Q., MAY, R. C., IBRAHIM, J. G., CHU, H. and COLE, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Stat. Med.* **33** 4560–4576. MR3267382 https://doi.org/10.1002/sim.6242

CURRIE, I. D. and DURBAN, M. (2002). Flexible smoothing with P-splines: A unified approach. *Stat. Model.* **2** 333–349. MR1951589 https://doi.org/10.1191/1471082x02st039ob

DRIKVANDI, R., KHODADADI, A. and VERBEKE, G. (2012). Testing variance components in balanced linear growth curve models. *J. Appl. Stat.* **39** 563–572. MR2880434 https://doi.org/10.1080/02664763.2011.603294

DRIKVANDI, R. and NOORIAN, S. (2019). Testing random effects in linear mixed-effects models with serially correlated errors. *Biom. J.* **61** 802–812. MR3982417 https://doi.org/10.1002/bimj.201700203

DRIKVANDI, R., VERBEKE, G., KHODADADI, A. and PARTOVINIA, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics* **14** 144–159.

DRIKVANDI, R., VERBEKE, G. and MOLENBERGHS, G. (2024). Supplement to "A framework for analysing longitudinal data involving time-varying covariates." https://doi.org/10.1214/23-AOAS1851SUPP

FERRER, E. and MCARDLE, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Struct. Equ. Model.* **10** 493–524. MR2011191 https://doi.org/10.1207/S15328007SEM1004_1

FIEUWS, S. and VERBEKE, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62** 424–431. MR2227490 https://doi.org/10.1111/j.1541-0420.2006.00507.x

GHOSH, P. and TU, W. (2009). Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. *J. Amer. Statist. Assoc.* **104** 474–485. MR2751432 https://doi.org/10.1198/jasa.2009.0013

Key words and phrases. AIDS cohort study, joint mixed model, longitudinal data, temporal association, time-varying covariate.

GUEORGUIEVA, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat. Model.* **1** 177–193.

HERNÁN, M. A., BRUMBACK, B. A. and ROBINS, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Stat. Med.* **21** 1689–1709. https://doi.org/10.1002/sim.1144

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

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

KIM, S. and ALBERT, P. S. (2016). A class of joint models for multivariate longitudinal measurements and a binary event. *Biometrics* **72** 917–925. MR3545684 https://doi.org/10.1111/biom.12463

KÜRÜM, E., JESKE, D. R., BEHRENDT, C. E. and LEE, P. (2018). A copula model for joint modeling of longitudinal and time-invariant mixed outcomes. *Stat. Med.* **37** 3931–3943. MR3873692 https://doi.org/10.1002/sim.7855

LI, H., ZHANG, Y., CARROLL, R. J., KOZEY KEADLE, S., SAMPSON, J. N. and MATTHEWS, C. E. (2017). A joint modeling and estimation method for multivariate longitudinal data with mixed types of responses to analyze physical activity data generated by accelerometers. *Stat. Med.* **36** 4028–4040. MR3713646 https://doi.org/10.1002/sim.7401

LIN, T.-I. and WANG, W.-L. (2013). Multivariate skew-normal linear mixed models for multi-outcome longitudinal data. *Stat. Model.* **13** 199–221. MR3179524 https://doi.org/10.1177/1471082X13480283

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398. https://doi.org/10.1093/biostatistics/5.3.381

PROUDFOOT, J., FAIG, W., NATARAJAN, L. and XU, R. (2018). A joint marginal-conditional model for multivariate longitudinal data. *Stat. Med.* **37** 813–828. MR3760451 https://doi.org/10.1002/sim.7552

RAO, K., DRIKVANDI, R. and SAVILLE, B. (2019). Permutation and Bayesian tests for testing random effects in linear mixed-effects models. *Stat. Med.* **38** 5034–5047. MR4022844 https://doi.org/10.1002/sim.8350

ROY, J., ALDERSON, D., HOGAN, J. W. and TASHIMA, K. T. (2006). Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women. *J. Amer. Statist. Assoc.* **101** 424–434. MR2256164 https://doi.org/10.1198/016214505000001203

ROY, J. and LIN, X. (2005). Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics* **61** 837–846. MR2196173 https://doi.org/10.1111/j.1541-0420.2005.00340.x

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 https://doi.org/10.1017/CBO9780511755453

STRAM, D. O. and LEE, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

SY, J. P., TAYLOR, J. M. and CUMBERLAND, W. G. (1997). A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* **53** 542–555.

THIÉBAUT, R., JACQMIN-GADDA, H., BABIKER, A., COMMENGES, D. and COLLABORATION, T. C. (2005). Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat. Med.* **24** 65–82. MR2134496 https://doi.org/10.1002/sim.1923

VERBEKE, G. and MOLENBERGHS, G. (2009). *Linear Mixed Models for Longitudinal Data*. *Springer Series in Statistics*. Springer, New York. MR2723365

WAND, M. P. (2003). Smoothing and mixed models. *Comput. Statist.* **18** 223–249.

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

XIANG, D., QIU, P. and PU, X. (2013). Nonparametric regression analysis of multivariate longitudinal data. *Statist. Sinica* **23** 769–789. MR3086655

ZHAO, L., CHEN, T., NOVITSKY, V. and WANG, R. (2021). Joint penalized spline modeling of multivariate longitudinal data, with application to HIV-1 RNA load levels and CD4 cell counts. *Biometrics* **77** 1061–1074. MR4320678 https://doi.org/10.1111/biom.13339

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1642–1667
https://doi.org/10.1214/23-AOAS1852

VARIANCE AS A PREDICTOR OF HEALTH OUTCOMES: SUBJECT-LEVEL TRAJECTORIES AND VARIABILITY OF SEX HORMONES TO PREDICT BODY FAT CHANGES IN PERI- AND POSTMENOPAUSAL WOMEN

BY IRENA CHEN[1,a], ZHENKE WU[1,b], SIOBÁN D. HARLOW[2,d], CARRIE A. KARVONEN-GUTIERREZ[2,e], MICHELLE M. HOOD[2,f] AND MICHAEL R. ELLIOTT[1,c]

[1]*Department of Biostatistics, University of Michigan,* [a]*irena@umich.edu,* [b]*zhenkewu@umich.edu,* [c]*mrelliot@umich.edu*
[2]*Department of Epidemiology, University of Michigan,* [d]*harlow@umich.edu,* [e]*ckarvone@umich.edu,* [f]*mmhood@umich.edu*

Longitudinal biomarker data and cross-sectional outcomes are routinely collected in modern epidemiology studies, often with the goal of informing tailored early intervention decisions. For example, hormones, such as estradiol (E2) and follicle-stimulating hormone (FSH), may predict changes in womens' health during the midlife. Most existing methods focus on constructing predictors from mean marker trajectories. However, subject-level biomarker variability may also provide critical information about disease risks and health outcomes. Current literature does not provide statistical models to investigate such relationships with valid uncertainty quantification. In this paper we develop a fully Bayesian joint model that estimates subject-level means, variances, and covariances of multiple longitudinal biomarkers and uses these as predictors to evaluate their respective associations with a cross-sectional health outcome. Simulations demonstrate excellent recovery of true model parameters. The proposed method provides less biased and more efficient estimates, relative to alternative approaches that either ignore subject-level differences in variances or perform two-stage estimation where estimated marker variances are treated as observed. Empowered by the model, analyses of women's health data reveal, for the first time, that larger variability of E2 was associated with slower increases in waist circumference across the menopausal transition.

REFERENCES

BARTLETT, J. W. and KEOGH, R. H. (2018). Bayesian correction for covariate measurement error: A frequentist evaluation and comparison with regression calibration. *Stat. Methods Med. Res.* **27** 1695–1708. MR3803260 https://doi.org/10.1177/0962280216667764

BÜRKNER, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *J. Stat. Softw.* **80** 1–28.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

CARR, M. C. (2003). The emergence of the metabolic syndrome with menopause. *J. Clin. Endocrinol. Metab.* **88** 2404–2411.

CARROLL, R. J. (2003). Variances are not always nuisance parameters. *Biometrics* **59** 211–220. MR1987387 https://doi.org/10.1111/1541-0420.t01-1-00027

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A modern perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

CHARANDABI, S. M., REZAEI, N., HAKIMI, S., MONTAZERI, A., TAHERI, S., TAGHINEJAD, H. and SAYEHMIRI, K. (2015). Quality of life of postmenopausal women and their spouses: A community-based study. *Iran Red Crescent. Med. J.* **17** e21599. https://doi.org/10.5812/ircmj.21599

Key words and phrases. Estradiol, follicle-stimulating hormone, Hamiltonian Monte Carlo, joint models, menopause, subject-level variability, Study of Women's Health Across the Nation (SWAN), variance component priors.

CHEN, I., WU, Z., HARLOW, S. D, KARVONEN-GUTIERREZ, C. A, HOOD, M. M and ELLIOTT, M. R (2024). Supplement to "Variance as a predictor of health outcomes: Subject-level trajectories and variability of sex hormones to predict body fat changes in peri- and postmenopausal women." https://doi.org/10.1214/23-AOAS1852SUPPA, https://doi.org/10.1214/23-AOAS1852SUPPB

CHI, Y.-Y. and IBRAHIM, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62** 432–445. MR2227491 https://doi.org/10.1111/j.1541-0420.2005.00448.x

COLLELUORI, G., CHEN, R., NAPOLI, N., AGUIRRE, L. E., QUALLS, C., VILLAREAL, D. T. and ARMAMENTO-VILLAREAL, R. (2018). Fat mass follows a U-shaped distribution based on estradiol levels in postmenopausal women. *Front. Endocrinol.* **9** 315.

DARSINI, D., HAMIDAH, H., NOTOBROTO, H. B. and CAHYONO, E. A. (2020). Health risks associated with high waist circumference: A systematic review. *J. Public Health Res.* **9** 1811. https://doi.org/10.4081/jphr.2020.1811

ELLIOTT, M. R., SAMMEL, M. D. and FAUL, J. (2012). Associations between variability of risk factors and health outcomes in longitudinal studies. *Stat. Med.* **31** 2745–2756. MR2972319 https://doi.org/10.1002/sim.5370

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

GHOSH, R. P., MALLICK, B. and POURAHMADI, M. (2021). Bayesian estimation of correlation matrices of longitudinal data. *Bayesian Anal.* **16** 1039–1058. MR4303878 https://doi.org/10.1214/20-BA1237

GORDON, J. L., RUBINOW, D. R., EISENLOHR-MOUL, T. A., LESERMAN, J. and GIRDLER, S. S. (2016). Estradiol variability, stressful life events, and the emergence of depressive symptomatology during the menopausal transition. *Menopause* **23** 257–266.

GOULD, A. L., BOYE, M. E., CROWTHER, M. J., IBRAHIM, J. G., QUARTEY, G., MICALLEF, S. and BOIS, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: Current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat. Med.* **34** 2181–2195. MR3354152 https://doi.org/10.1002/sim.6141

GOURLAY, M. L., SPECKER, B. L., LI, C., HAMMETT-STABLER, C. A., RENNER, J. B. and RUBIN, J. E. (2012). Follicle-stimulating hormone is independently associated with lean mass but not BMD in younger postmenopausal women. *Bone* **50** 311–316. https://doi.org/10.1016/j.bone.2011.11.001

GREENDALE, G. A., HAN, W., FINKELSTEIN, J. S., BURNETT-BOWIE, S.-A. M., HUANG, M., MARTIN, D. and KARLAMANGLA, A. S. (2021). Changes in regional fat distribution and anthropometric measures across the menopause transition. *J. Clin. Endocrinol. Metab.* **106** 2520–2534.

GREENDALE, G. A., STERNFELD, B., HUANG, M., HAN, W., KARVONEN-GUTIERREZ, C., RUPPERT, K., CAULEY, J. A., FINKELSTEIN, J. S., JIANG, S.-F. et al. (2019). Changes in body composition and weight during the menopause transition. *JCI Insight* **4** e124865.

GRILICHES, Z. and INTRILIGATOR, M. D. (1987). Handbook of Econometrics 25, 1465–1514. North Holland.

HARLOW, S. D., LIN, X. and HO, M. J. (2000). Analysis of menstrual diary data across the reproductive life span applicability of the bipartite model approach and the importance of within-woman variance. *J. Clin. Epidemiol.* **53** 722–733.

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480. https://doi.org/10.1093/biostatistics/1.4.465

HUANG, X., ELLIOTT, M. R. and HARLOW, S. D. (2014). Modelling menstrual cycle length and variability at the approach of menopause by using hierarchical change point models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 445–466. MR3238161 https://doi.org/10.1111/rssc.12044

IBRAHIM, J. G., CHU, H. and CHEN, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol.* **28** 2796–2801.

JIANG, B., ELLIOTT, M. R., SAMMEL, M. D. and WANG, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics* **71** 487–497. MR3366253 https://doi.org/10.1111/biom.12284

KARVONEN-GUTIERREZ, C. and HARLOW, S. D. (2017). Menopause and midlife health changes. In *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. (J. B. Halter, J. G. Ouslander, S. Studenski, K. P. High, S. Asthana, M. A. Supiano and C. Ritchie, eds.) McGraw-Hill Education, New York.

KOHRT, W. M. and WIERMAN, M. E. (2017). Preventing fat gain by blocking follicle-stimulating hormone. *N. Engl. J. Med.* **377** 293–295. https://doi.org/10.1056/NEJMcibr1704542

LEWANDOWSKI, D., KUROWICKA, D. and JOE, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivariate Anal.* **100** 1989–2001. MR2543081 https://doi.org/10.1016/j.jmva.2009.04.008

LIU, P., JI, Y., YUEN, T., RENDINA-RUEDY, E., DEMAMBRO, V. E., DHAWAN, S., ABU-AMER, W., IZADMEHR, S., ZHOU, B. et al. (2017). Blocking FSH induces thermogenic adipose tissue and reduces body fat. *Nature* **546** 107–112.

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** 138. https://doi.org/10.1186/s12874-018-0592-9

OGBURN, E. L., RUDOLPH, K. E., MORELLO-FROSCH, R., KHAN, A. and CASEY, J. A. (2021). A warning about using predicted values from regression models for epidemiologic inquiry. *Amer. J. Epidemiol.* **190** 1142–1147. https://doi.org/10.1093/aje/kwaa282

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

PARK, S. K., HARLOW, S. D., ZHENG, H., KARVONEN-GUTIERREZ, C., THURSTON, R. C., RUPPERT, K., JANSSEN, I. and RANDOLPH, J. F. (2017). Association between changes in oestradiol and follicle-stimulating hormone levels during the menopausal transition and risk of diabetes. *Diabet. Med.: J. Brit. Diabet. Assoc.* **34** 531–538.

PETTEE GABRIEL, K., STERNFELD, B., COLVIN, A., STEWART, A., STROTMEYER, E. S., CAULEY, J. A., DUGAN, S. and KARVONEN-GUTIERREZ, C. (2017). Physical activity trajectories during midlife and subsequent risk of physical functioning decline in late mid-life: The study of women's health across the nation (SWAN). *Prev. Med.* **105** 287–294.

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

RANDOLPH, J. F., ZHENG, H., SOWERS, M. R., CRANDALL, C., CRAWFORD, S., GOLD, E. B. and VUGA, M. (2011). Change in follicle-stimulating hormone and estradiol across the menopausal transition: Effect of age at the final menstrual period **96** 746–754.

RANDOLPH, J. F. JR., SOWERS, M., BONDARENKO, I. V., HARLOW, S. D., LUBORSKY, J. L. and LITTLE, R. J. (2004). Change in estradiol and follicle-stimulating hormone across the early menopausal transition: Effects of ethnicity and age. *J. Clin. Endocrinol. Metab.* **89** 1555–1561.

REES, M., BITZER, J., CANO, A., CEAUSU, I., CHEDRAUI, P., DURMUSOGLU, F., ERKKOLA, R., GEUKES, M., GODFREY, A. et al. (2021). Global consensus recommendations on menopause in the workplace: A European menopause and andropause society (EMAS) position statement. *Maturitas* **151** 55–62.

RICHARDSON, S. and GILKS, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Amer. J. Epidemiol.* **138** 430–442. https://doi.org/10.1093/oxfordjournals.aje.a116875

ROBERT, C. P. and CASELLA, G. (2010). Monte Carlo integration. In *Introducing Monte Carlo Methods with R* (C. Robert and G. Casella, eds.). *Use R* 61–88. Springer.

ROSS, R., NEELAND, I. J., YAMASHITA, S., SHAI, I., SEIDELL, J., MAGNI, P., SANTOS, R. D., ARSENAULT, B., CUEVAS, A. et al. (2020). Waist circumference as a vital sign in clinical practice: A consensus statement from the IAS and ICCR working group on visceral obesity. *Nat. Rev. Endocrinol.* **16** 177–189.

SAMMEL, M., WANG, Y., RATCLIIE, S., FREEMAN, E. and PROPERT, K. (2001). Models for within-subject heterogeneity as predictors for disease. In *Proceedings of the Annual Meeting of the American Statistical Association*.

SOWERS, M., CRAWFORD, S. L., STERNFELD, B., MORGANSTEIN, D., GOLD, E. B., GREENDALE, G. A., EVANS, D., NEER, R., MATTHEWS, K. et al. (2000). SWAN: A multicenter, multiethnic, community-based cohort study of women and the menopausal transition. In *Menopause*: *Biology and Pathology* (R. A. Lobo, J. Kelsey and R. Marcus, eds.) 175–188. Academic Press, San Diego.

SOWERS, M., ZHENG, H., TOMEY, K., KARVONEN-GUTIERREZ, C., JANNAUSCH, M., LI, X., YOSEF, M. and SYMONS, J. (2007). 6-year changes in body composition in women at mid-life: Ovarian and chronological aging. *J. Clin. Endocrinol. Metab.* **92** 895–901.

SPONTON, C. H. and KAJIMURA, S. (2017). Burning fat and building bone by FSH blockade. *Cell Metab.* **26** 285–287. https://doi.org/10.1016/j.cmet.2017.07.018

STAN DEVELOPMENT TEAM (2020). RStan: The R interface to Stan. R package version 2.21.2.

STAN DEVELOPMENT TEAM (2023). 6.1 Bayesian measurement error model | Stan User's Guide.

STEVENS, J., CAI, J., EVENSON, K. R. and THOMAS, R. (2002). Fitness and fatness as predictors of mortality from all causes and from cardiovascular disease in men and women in the lipid research clinics study. *Amer. J. Epidemiol.* **156** 832–841.

U. S. CENSUS BUREAU 2017 National Population Projections Tables: Main Series. Section: Government.

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. MR4298989 https://doi.org/10.1214/20-ba1221

WANG, J., LUO, S. and LI, L. (2017). Dynamic prediction for multiple repeated measures and event time data: An application to Parkinson's disease. *Ann. Appl. Stat.* **11** 1787–1809. MR3709578 https://doi.org/10.1214/17-AOAS1059

WANG, S., MCCORMICK, T. H. and LEEK, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proc. Natl. Acad. Sci. USA* **117** 30266–30275. MR4263300 https://doi.org/10.1073/pnas.2001238117

YOUNG, H. A. and BENTON, D. (2018). Heart-rate variability: A biomarker to study the influence of nutrition on physiological and psychological health? *Behav. Pharmacol.* **29** 140–151.

ZAIDI, M., LIZNEVA, D., KIM, S.-M., SUN, L., IQBAL, J., NEW, M. I., ROSEN, C. J. and YUEN, T. (2018). FSH, bone mass, body fat, and biological aging. *Endocrinology* **159** 3503–3514.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1668–1685
https://doi.org/10.1214/23-AOAS1853

FUNCTIONAL CONCURRENT REGRESSION WITH COMPOSITIONAL COVARIATES AND ITS APPLICATION TO THE TIME-VARYING EFFECT OF CAUSES OF DEATH ON HUMAN LONGEVITY

BY EMANUELE GIOVANNI DEPAOLI[1,a], MARCO STEFANUCCI[2,c] AND STEFANO MAZZUCO[1,b]

[1]*Department of Statistical Sciences, University of Padova,* [a]*depaoli@stat.unipd.it,* [b]*stefano.mazzuco@unipd.it*
[2]*Department of Economics and Finance, University of Rome Tor Vergata,* [c]*marco.stefanucci@uniroma2.it*

Multivariate functional data that are cross-sectionally compositional data are attracting increasing interest in the statistical modeling literature, a major example being trajectories over time of compositions derived from cause-specific mortality rates. In this work we develop a novel functional concurrent regression model in which independent variables are functional compositions. This allows us to investigate the relationship over time between life expectancy at birth and compositions derived from cause-specific mortality rates of four distinct age classes, namely, zero to four, five to 39, 40–64 and 65+ in 25 countries. A penalized approach is developed to estimate the regression coefficients and select the relevant variables. Then an efficient computational strategy, based on an augmented Lagrangian algorithm, is derived to solve the resulting optimization problem. The good performances of the model in predicting the response function and estimating the unknown functional coefficients are shown in a simulation study. The results on real data confirm the important role of neoplasms and cardiovascular diseases in determining life expectancy emerged in other studies and reveal several other contributions not yet observed.

REFERENCES

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press, Caldwell.

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

BARBIERI, M., WILMOTH, J. R., SHKOLNIKOV, V. M., GLEI, D., JASILIONIS, D., JDANOV, D., BOE, C., RIFFE, T., GRIGORIEV, P. et al. (2015). Data resource profile: The human mortality database (HMD). *Int. J. Epidemiol.* **44** 1549–1556. https://doi.org/10.1093/ije/dyv105

BERGERON-BOUCHER, M.-P., ABURTO, J. M. and VAN RAALTE, A. (2020). Diversification in causes of death in low-mortality countries: Emerging patterns and implications. *BMJ Glob. Health* **5**. https://doi.org/10.1136/bmjgh-2020-002414

BERTSEKAS, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. *Computer Science and Applied Mathematics*. Academic Press, San Diego. MR0690767

BLACHIER, M., LELEU, H., PECK-RADOSAVLJEVIC, M., VALLA, D.-C. and ROUDOT-THORAVAL, F. (2013). The burden of liver disease in Europe: A review of available epidemiological data. *J. Hepatol.* **58** 593–608. https://doi.org/10.1016/j.jhep.2012.12.005

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers* **3** 1–122. Foundations and Trends® in Machine Learning.

CANUDAS-ROMO, V. (2008). The modal age at death and the shifting mortality hypothesis. *Demogr. Res.* **19** 1179–1204.

CANUDAS-ROMO, V. (2010). Three measures of longevity: Time trends and record values. *Demography* **47** 299–312. https://doi.org/10.1353/dem.0.0098

CANUDAS-ROMO, V., ADAIR, T. and MAZZUCO, S. (2020). Reflection on modern methods: Cause of death decomposition of cohort survival comparisons. *Int. J. Epidemiol.* **49** 1712–1718. https://doi.org/10.1093/ije/dyz276

Key words and phrases. Mortality by cause, life expectancy, functional data analysis, compositional data analysis, sparsity.

COENDERS, G. and PAWLOWSKY-GLAHN, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT* **44** 201–220. MR4121257 https://doi.org/10.2436/20.8080.02.100

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

DEPAOLI, E. G., STEFANUCCI, M. and MAZZUCO, S. (2024). Supplement to "Functional concurrent regression with compositional covariates and its application to the time-varying effect of causes of death on human longevity." https://doi.org/10.1214/23-AOAS1853SUPP

FERALDI, A. and ZARRULLI, V. (2022). Patterns in age and cause of death contribution to the sex gap in life expectancy: A comparison among ten countries. *Genus* **78**. https://doi.org/10.1186/s41118-022-00171-9

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

HUMAN CAUSE-OF-DEATH DATABASE. French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany). Available at www.causeofdeath.org.

HUMAN MORTALITY DATABASE. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research. (Germany). Available at www.mortality.org or www.humanmortality.de.

JANI, C., MARSHALL, D. C., SINGH, H., GOODALL, R., SHALHOUB, J., OMARI, O. A., SALCICCIOLI, J. D. and THOMSON, C. C. (2021). Lung cancer mortality in Europe and the USA between 2000 and 2017: An observational analysis. *ERJ Open Res*. **7**. https://doi.org/10.1183/23120541.00311-2021

JASILIONIS, D., VAN RAALTE, A. A., KLÜSENER, S. and GRIGORIEV, P. (2023). The underwhelming German life expectancy. *Eur. J. Epidemiol*. https://doi.org/10.1007/s10654-023-00995-5

KJÆRGAARD, S., ERGEMEN, Y. E., KALLESTRUP-LAMB, M., OEPPEN, J. and LINDAHL-JACOBSEN, R. (2019). Forecasting causes of death by using compositional data analysis: The case of cancer deaths. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **68** 1351–1370. MR4022816 https://doi.org/10.1111/rssc.12357

LEWER, D., JAYATUNGA, W., ALDRIDGE, R. W., EDGE, C., MARMOT, M., STORY, A. and HAYWARD, A. (2020). Premature mortality attributable to socioeconomic inequality in England between 2003 and 2018: An observational study. *Lancet Public Health* **5** e33–e41. https://doi.org/10.1016/S2468-2667(19)30219-1

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917 https://doi.org/10.1093/biomet/asu031

MEHTA, N. K., ABRAMS, L. R. and MYRSKYLÄ, M. (2020). US life expectancy stalls due to cardiovascular disease, not drug deaths. *Proc. Natl. Acad. Sci. USA* **117** 6998–7000. https://doi.org/10.1073/pnas.1920391117

MESLÉ, F. (2004). Mortality in Central and Eastern Europe: Long-term trends and recent upturns. *Demogr. Res*. **2** 45–70.

OEPPEN, J. (2008). Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data, Barcelona, Spain. Paper presented at the European Population Conference 2008.

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

REMUND, A., CAMARDA, C. G. and RIFFE, T. (2018). A cause-of-death decomposition of young adult excess mortality. *Demography* **55** 957–978. https://doi.org/10.1007/s13524-018-0680-9

SHI, P., ZHANG, A. and LI, H. (2016). Regression analysis for microbiome compositional data. *Ann. Appl. Stat*. **10** 1019–1040. MR3528370 https://doi.org/10.1214/16-AOAS928

STEFANUCCI, M. and MAZZUCO, S. (2022). Analysing cause-specific mortality trends using compositional functional data analysis. *J. Roy. Statist. Soc. Ser. A* **185** 61–83. MR4384297 https://doi.org/10.1111/rssa.12715

SUN, Z., XU, W., CONG, X., LI, G. and CHEN, K. (2020). Log-contrast regression with functional compositional predictors: Linking preterm infants' gut microbiome trajectories to neurobehavioral outcome. *Ann. Appl. Stat*. **14** 1535–1556. MR4152145 https://doi.org/10.1214/20-AOAS1357

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

VAN DE WATER H. P. A. (1997). Health expectancy and the problem of substitute morbidity. *Philos. Trans. R. Soc. Lond. B, Biol. Sci*. **352** 1819–1827.

VAUPEL, J. W. and CANUDAS-ROMO, V. (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography* **40** 201–216.

WHO MORTALITY DATABASE. World Health Organization. Available at www.who.int/data/data-collection-tools/who-mortality-database.

WOOLF, S. H. and SCHOOMAKER, H. (2019). Life expectancy and mortality rates in the United States, 1959–2017. *JAMA* **322** 1996–2016. https://doi.org/10.1001/jama.2019.16932

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

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1686–1708
https://doi.org/10.1214/23-AOAS1854

HOW ARE PRELAUNCH ONLINE MOVIE REVIEWS RELATED TO BOX OFFICE REVENUES?

BY TIANYU GUAN[1,a], JASON HO[2,b], ROBERT KRIDER[2,c], JIGUO CAO[3,d] AND ANDREW FOGG[4,e]

[1]*Department of Mathematics and Statistics, York University,* [a]*tianyug1988@gmail.com*
[2]*Beedie School of Business, Simon Fraser University,* [b]*jason_ho_3@sfu.ca,* [c]*robert_krider@sfu.ca*
[3]*Department of Statistics and Actuarial Science, Simon Fraser University,* [d]*jiguo_cao@sfu.ca*
[4]*Roku, Inc.,* [e]*afogg@roku.com*

This paper studies the dynamic patterns of the prelaunch online movie reviews, or movie electronic word-of-mouth (eWOM), over time and investigates their relations to the subsequent box office revenues. The volume and valence of prelaunch eWOM have been shown to be early indicators of strong or weak box office. The time patterns of prelaunch eWOM evolution, which are essentially functional data, on the other hand, tend to be overlooked. We apply the functional principal component analysis, a dimension reduction technique in functional data analysis, to analyze the dynamic patterns of various quantile trajectories of the movie eWOM, instead of directly studying the whole eWOM functional data. The functional principal component (FPC) scores of quantile trajectories at various quantile levels are used to predict the box office revenues. We use the sparse group lasso method to select the quantile levels and individual FPC scores that make significant contributions to the prediction of box office revenues. The results show that compared with other measures, such as valence and variance, the top-end quantiles would be a better measure in capturing the relations between the prelaunch product ratings time pattern and launch sales.

REFERENCES

BABIĆ ROSARIO, A., DE VALCK, K. and SOTGIU, F. (2020). Conceptualizing the electronic word-of-mouth process: What we know and need to know about eWOM creation, exposure, and evaluation. *J. Acad. Mark. Sci.* **48** 422–448.

BABIĆ ROSARIO, A., SOTGIU, F., DE VALCK, K. and BIJMOLT, T. H. A. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *J. Mark. Res.* **53** 297–318.

BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343 https://doi.org/10.1214/07-AOS516

BIKHCHANDANI, S., HIRSHLEIFER, D. and WELCH, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.* **100** 992–1026.

CARDOT, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Stat.* **12** 503–538. MR1785396 https://doi.org/10.1080/10485250008832820

CHINTAGUNTA, P. K., GOPINATH, S. and VENKATARAMAN, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Mark. Sci.* **29** 944–957.

CLEMONS, E. K., GAO, G. and HITT, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *J. Manage Inf. Syst.* **23** 149–171.

DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. MR0650934 https://doi.org/10.1016/0047-259X(82)90088-4

DHAR, V. and CHANG, E. A. (2009). Does chatter matter? The impact of user-generated content on music sales. *J. Interact. Mark.* **23** 300–307.

Key words and phrases. Electronic word of mouth, functional data analysis, functional principal component analysis, quantile functions, variable selection.

ELIASHBERG, J., HEGIE, Q., HO, J., HUISMAN, D., MILLER, S. J., SWAMI, S., WIERENGA, C. B. and WIERENGA, B. (2009). Demand-driven scheduling of movies in a multiplex. *Int. J. Res. Mark.* **26** 75–88.

FOUTZ, N. Z. and JANK, W. (2010). Research note—Prerelease demand forecasting for motion pictures using functional shape analysis of virtual stock markets. *Mark. Sci.* **29** 568–579.

GELPER, S., PERES, R. and ELIASHBERG, J. (2018). Talk bursts: The role of spikes in prerelease word-of-mouth dynamics. *J. Mark. Res.* **55** 801–817.

GHOSAL, R., VARMA, V. R., VOLFSON, D., HILLEL, I., URBANEK, J., HAUSDORFF, J. M., WATTS, A. and ZIPUNNIKOV, V. (2023). Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's Disease. *Biostatistics* **24** 539–561. MR4615240 https://doi.org/10.1093/biostatistics/kxab041

GIL, R. and HARTMANN, W. R. (2009). Empirical analysis of metering price discrimination: Evidence from concession sales at movie theaters. *Mark. Sci.* **28** 1046–1062.

GILCHRIST, W. (2000). *Statistical Modelling with Quantile Functions*. CRC Press/CRC, Boca Raton, FL.

GUAN, T., HO, J., KRIDER, R., CAO, J. and FOGG, A. (2024). Supplement to "How are PreLaunch online movie reviews related to box office revenues?." https://doi.org/10.1214/23-AOAS1854SUPP

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 109–126. MR2212577 https://doi.org/10.1111/j.1467-9868.2005.00535.x

HO, J. Y. C., LIANG, Y., WEINBERG, C. B. and YAN, J. (2018). An empirical study of uniform and differential pricing in the movie theatrical market. *J. Mark. Res.* **55** 414–431.

HOUSTON, M. B., KUPFER, A. K., HENNIG-THURAU, T. and SPANN, M. (2018). Pre-release consumer buzz. *J. Acad. Mark. Sci.* **46** 338–360.

HU, N., PAVLOU, P. A. and ZHANG, J. (2009). Overcoming the J-shaped distribution of product reviews. *Commun. ACM* **52** 144–147.

HU, N., PAVLOU, P. A. and ZHANG, J. (2017). On self-selection biases in online product reviews. *MIS Q.* **41** 449–471.

JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. MR1789811 https://doi.org/10.1093/biomet/87.3.587

VERMA, S. and YADAV, N. (2021). Past, present, and future of electronic word of mouth (eWOM). *J. Interact. Mark.* **53** 111–128.

LATANE, B. and WOLF, S. (1981). The social impact of majorities and minorities. *Psychol. Rev.* **88** 438–453.

LI, X. and HITT, L. M. (2008). Self-selection and information role of online product reviews. *Inf. Syst. Res.* **19** 456–474.

LI, Y., WANG, N. and CARROLL, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108** 1284–1294. MR3174708 https://doi.org/10.1080/01621459.2013.788980

LIN, Z., WANG, L. and CAO, J. (2016). Interpretable functional principal component analysis. *Biometrics* **72** 846–854. MR3545677 https://doi.org/10.1111/biom.12457

LIU, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Mark.* **70** 74–89.

MUDAMBI, S. M. and SCHUFF, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Q.* **34** 185–200.

NIE, Y. and CAO, J. (2020). Sparse functional principal component analysis in a new regression framework. *Comput. Statist. Data Anal.* **152** 107016, 15. MR4114152 https://doi.org/10.1016/j.csda.2020.107016

NIE, Y., WANG, L., LIU, B. and CAO, J. (2018). Supervised functional principal component analysis. *Stat. Comput.* **28** 713–723. MR3761351 https://doi.org/10.1007/s11222-017-9758-2

NIE, Y., YANG, Y., WANG, L. and CAO, J. (2022). Recovering the underlying trajectory from sparse and irregular longitudinal data. *Canad. J. Statist.* **50** 122–141. MR4389173 https://doi.org/10.1002/cjs.11677

PAUWELS, K., AKSEHIRLI, Z. and LACKMAN, A. (2016). Like the ad or the brand? Marketing stimulates different electronic word-of-mouth content to drive online and offline performance. *Int. J. Res. Mark.* **33** 639–655.

PURNAWIRAWAN, N., EISEND, M., DE PELSMACKER, P. and DENS, N. (2015). A meta-analytic investigation of the role of valence in online reviews. *J. Interact. Mark.* **31** 17–27.

QAHRI-SAREMI, H. and MONTAZEMI, A. R. (2019). Factors affecting the adoption of an electronic word of mouth message: A meta-analysis. *J. Manage Inf. Syst.* **36** 969–1001.

RAMSAY, J. O., HOOKER, G. and GRAVES, S. (2009). *Functional Data Analysis with R and Matlab*. Springer, New York.

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

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

SANG, P., BEGEN, M. A. and CAO, J. (2021). Appointment scheduling with a quantile objective. *Comput. Oper. Res.* **132** Paper No. 105295, 20. MR4255393 https://doi.org/10.1016/j.cor.2021.105295

SANG, P., WANG, L. and CAO, J. (2017). Parametric functional principal component analysis. *Biometrics* **73** 802–810. MR3713114 https://doi.org/10.1111/biom.12641

SHI, H., DONG, J., WANG, L. and CAO, J. (2021). Functional principal component analysis for longitudinal data with informative dropout. *Stat. Med.* **40** 712–724. MR4198440 https://doi.org/10.1002/sim.8798

SHI, H., YANG, Y., WANG, L., MA, D., BEG, M. F., PEI, J. and CAO, J. (2022). Two-dimensional functional principal component analysis for image feature extraction. *J. Comput. Graph. Statist.* **31** 1127–1140. MR4513375 https://doi.org/10.1080/10618600.2022.2035738

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250

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

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Review of functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

XIONG, G. and BHARADWAJ, S. (2014). Prerelease buzz evolution patterns and new product performance. *Mark. Sci.* **33** 401–421.

YANG, H., BALADANDAYUTHAPANI, V., RAO, A. U. K. and MORRIS, J. S. (2020). Quantile function on scalar regression analysis for distributional data. *J. Amer. Statist. Assoc.* **115** 90–106. MR4078447 https://doi.org/10.1080/01621459.2019.1609969

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

YOU, Y., VADAKKEPATT, G. G. and JOSHI, A. M. (2015). A meta-analysis of electronic word-of-mouth elasticity. *J. Mark.* **79** 19–39.

ZHANG, J.-T. and CHEN, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35** 1052–1079. MR2341698 https://doi.org/10.1214/009053606000001505

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1709–1728
https://doi.org/10.1214/23-AOAS1855

A HIERARCHICAL SPLINE MODEL FOR CORRECTING AND HINDCASTING TEMPERATURE DATA

BY THEODOROS ECONOMOU[1,a], CATRINA JOHNSON[2,b] AND ELIZABETH DYSON[2,c]

[1]*Climate and Atmospheric Research Centre, The Cyprus Institute,* [a]*t.economou@cyi.ac.cy*
[2]*UK Met Office,* [b]*catrina.johnson@metoffice.gov.uk,* [c]*elizabeth.dyson@metoffice.gov.uk*

Weather observations are important for a wide range of applications although they do pose statistical challenges, such as missing values, errors, flawed outliers and poor spatial and temporal coverage to name a few. A Bayesian hierarchical spline framework is presented here to deal with such challenges in temperature time series. Motivated by a real-life problem, the approach uses penalised splines, constructed hierarchically, to pool the data, along with a discrete mixture distribution to deal with outliers and publicly available global reanalysis data sets (climate model data) to integrate physically constrained information. Efficient Bayesian implementation is achieved using conditional conjugacy, which allows thorough model checking and uncertainty quantification. Fitting the model to daily maximum temperature illustrates its flexibility in capturing temporal structures, in pooling of the information and in outlier detection. The model is used to hindcast the time series 50 years into the past while maintaining uncertainty at reasonable levels.

REFERENCES

BARNETT, V. and LEWIS, T. (1994). *Outliers in Statistical Data*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, Chichester. MR1272911

BOÉ, J., TERRAY, L., HABETS, F. and MARTIN, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *Int. J. Climatol.* **27** 1643–1655.

DÉQUÉ, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Glob. Planet. Change* **57** 16–26.

ECONOMOU, T. (2023). Data, code and supplementary material for: A hierarchical spline model for correcting and hindcasting tempererature data.

ECONOMOU, T., JOHNSON, C. and DYSON, E. (2024). Supplement to "A hierarchical spline model for correcting and hindcasting temperature data." https://doi.org/10.1214/23-AOAS1855SUPP

ECONOMOU, T., LAZOGLOU, G., TZYRKALLI, A., CONSTANTINIDOU, K. and LELIEVELD, J. (2023). A data integration framework for spatial interpolation of temperature observations using climate model data. *PeerJ* **11** e14519. https://doi.org/10.7717/peerj.14519

FINK, D. (1997). A Compendium of Conjugate Priors Technical Report.

GARCÍA-ZATTERA, M. J., JARA, A. and KOMÁREK, A. (2016). A flexible AFT model for misclassified clustered interval-censored data. *Biometrics* **72** 473–483. MR3515774 https://doi.org/10.1111/biom.12424

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

GUDMUNDSSON, L., BREMNES, J. B., HAUGEN, J. E. and ENGEN-SKAUGEN, T. (2012). Technical note: Downscaling RCM precipitation to the station scale using statistical transformations—a comparison of methods. *Hydrol. Earth Syst. Sci.* **16** 3383–3390.

HAWKINS, D. M. (1980). *Identification of Outliers*. *Monographs on Applied Probability and Statistics*. CRC Press, London. MR0584791

HERSBACH, H., BELL, B., BERRISFORD, P., BIAVATI, G., HORÁNYI, A., MUÑOZ SABATER, J., NICOLAS, J., PEUBEY, C., RADU, R. et al. (1918). ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

HILL, D. J. and MINSKER, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environ. Model. Softw.* **25** 1014–1022.

HODGE, V. J. and AUSTIN, J. (2004). A survey of outlier detection methodologies. *Artif. Intell. Rev.* **22** 85–126.

Key words and phrases. Mixture model, MCMC, random effects, Bayesian splines, outliers.

HUNZIKER, S., GUBLER, S., CALLE, J., MORENO, I., ANDRADE, M., VELARDE, F., TICONA, L., CARRASCO, G., CASTELLÓN, Y. et al. (2017). Identifying, attributing, and overcoming common data quality issues of manned station observations. *Int. J. Climatol.* **37** 4131–4145.

JOBE, J. M. and POKOJOVY, M. (2015). A cluster-based outlier detection scheme for multivariate data. *J. Amer. Statist. Assoc.* **110** 1543–1551. MR3449053 https://doi.org/10.1080/01621459.2014.983231

KAMMANN, E. E. and WAND, M. P. (2003). Geoadditive models. *J. R. Stat. Soc., Ser. C* **52** 1–18. MR1963210 https://doi.org/10.1111/1467-9876.00385

LI, G. and JUNG, J. J. (2021). Dynamic graph embedding for outlier detection on multiple meteorological time series. *PLoS ONE* **16** 1–14.

LIU, H., WANG, B., SUN, X., LI, T., LIU, Q. and GUO, Y. (2014). DCSCS: A novel approach to improve data accuracy for low cost meteorological sensor networks. *Inf. Technol. J.* **13** 1640.

MA, L., GU, X. and WANG, B. (2017). Correction of outliers in temperature time series based on sliding window prediction in meteorological sensor network. *Information* **8**.

MAHMOOD, R., FOSTER, S. A. and LOGAN, D. (2006). The GeoProfile metadata, exposure of instruments, and measurement bias in climatic record revisited. *Int. J. Climatol.* **26** 1091–1124.

MENG, Z., ZHANG, S. C. and HUANG, Z. L. (2013). Outlier detection for observational data of automatic meteorological station based on least square support vector machine. In *Progress in Environmental Protection and Processing of Resource. Applied Mechanics and Materials* **295** 945–949. Trans Tech Publications Ltd.

NAYAK, D. and PERROS, H. (2020). Automated real-time anomaly detection of temperature sensors through machine-learning. *Int. J. Sens. Netw.* **34** 137–152.

PEDERSEN, E. J., MILLER, D. L., SIMPSON, G. L. and ROSS, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* **7** 341–360.

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.

RAMACHANDRA, B., DUTTON, B. and VATSAVAI, R. R. (2019). Anomalous cluster detection in spatiotemporal meteorological fields. *Stat. Anal. Data Min.* **12** 88–100. MR3928233 https://doi.org/10.1002/sam.11398

REUNANEN, N., RÄTY, T., JOKINEN, J. J., HOYT, T. and CULLER, D. (2020). Unsupervised online detection and prediction of outliers in streams of sensor data. *Int. J. Data Sci. Anal.* **9** 285–314.

RHODES, R. I., SHAFFREY, L. C. and GRAY, S. L. (2015). Can reanalyses represent extreme precipitation over England and Wales? *Q. J. R. Meteorol. Soc.* **141** 1114–1120.

ROBERTS, S., OSBORNE, M., EBDEN, M., REECE, S., GIBSON, N. and AIGRAIN, S. (2013). Gaussian processes for time-series modelling. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20110550. MR3005668 https://doi.org/10.1098/rsta.2011.0550

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

STEPANEK, P., ZAHRADNÍČEK, P. and FARDA, A. (2013). Experiences with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010. *Idojaras* **117** 123–141.

SUN, X., YAN, S., WANG, B., XIA, L., LIU, Q. and ZHANG, H. (2015). Air temperature error correction based on solar radiation in an economical meteorological wireless sensor network. *Sensors* **15** 18114–18139.

WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 95–114. MR1959095 https://doi.org/10.1111/1467-9868.00374

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734 https://doi.org/10.1111/j.1467-9868.2010.00749.x

WOOD, S. N. (2016). Just another Gibbs additive modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75** 1–15.

WOOD, S. N. (2017). *Generalized Additive Models*: *An Introduction with ß f R*, 2nd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3726911

WOOD, S. N., SCHEIPL, F. and FARAWAY, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Stat. Comput.* **23** 341–360. MR3041440 https://doi.org/10.1007/s11222-012-9314-z

WU, E., LIU, W. and CHAWLA, S. (2010). Spatio-temporal outlier detection in precipitation data. In *Knowledge Discovery from Sensor Data* (M. M. Gaber, R. R. Vatsavai, O. A. Omitaomu, J. Gama, N. V. Chawla and A. R. Ganguly, eds.) 115–133. Springer, Berlin.

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1729–1749
https://doi.org/10.1214/23-AOAS1856
This research was funded, in whole or in part, by Wellcome Trust, 204623/Z/16/Z.

SELECTING INVALID INSTRUMENTS TO IMPROVE MENDELIAN RANDOMIZATION WITH TWO-SAMPLE SUMMARY DATA

BY ASHISH PATEL[1,a], FRANCIS J. DITRAGLIA[2,c], VERENA ZUBER[3,d] AND STEPHEN BURGESS[1,b]

[1]*MRC Biostatistics Unit, University of Cambridge,* [a]*ashish.patel@mrc-bsu.cam.ac.uk,* [b]*sb452@medschl.cam.ac.uk*
[2]*Department of Economics, University of Oxford,* [c]*francis.ditraglia@economics.ox.ac.uk*
[3]*Department of Biostatistics and Epidemiology, Imperial College London,* [d]*verena.zuber@imperial.ac.uk*

Mendelian randomization (MR) is a widely-used method to estimate the causal relationship between a risk factor and disease. A fundamental part of any MR analysis is to choose appropriate genetic variants as instrumental variables. Genome-wide association studies often reveal that hundreds of genetic variants may be robustly associated with a risk factor, but in some situations investigators may have greater confidence in the instrument validity of only a smaller subset of variants. Nevertheless, the use of additional instruments may be optimal from the perspective of mean squared error, even if they are slightly invalid; a small bias in estimation may be a price worth paying for a larger reduction in variance. For this purpose we consider a method for "focused" instrument selection whereby genetic variants are selected to minimise the estimated asymptotic mean squared error of causal effect estimates. In a setting of many weak and locally invalid instruments, we propose a novel strategy to construct confidence intervals for postselection focused estimators that guards against the worst case loss in asymptotic coverage. In empirical applications to: (i) validate lipid drug targets and (ii) investigate vitamin D effects on a wide range of outcomes, our findings suggest that the optimal selection of instruments does not involve only a small number of biologically-justified instruments but also many potentially invalid instruments.

REFERENCES

ANDREWS, I. (2018). Valid two-step identification-robust confidence sets for GMM. *Rev. Econ. Stat.* **100** 337–348.

ARMSTRONG, T. B., KOLESÁR, M. and PLAGBORG-MØLLER, M. (2022). Robust empirical Bayes confidence intervals. *Econometrica* **90** 2567–2602. MR4524894 https://doi.org/10.3982/ecta18597

BARBARAWI, M., KHEIRI, B., ZAYED, Y., BARBARAWI, O., DHILLON, H., SWAID, B., YELANGI, A., SUNDUS, S., BACHUWA, G. et al. (2019). Vitamin D supplementation and cardiovascular disease risks in more than 83 000 individuals in 21 randomized clinical trials: A meta-analysis. *JAMA Cardiol.* **4** 765–776.

BERRY, D. J., VIMALESWARAN, K. S., WHITTAKER, J. C., HINGORANI, A. D. and HYPPÖNEN, E. (2012). Evaluation of genetic markers as instruments for Mendelian randomization studies on vitamin D. *PLoS ONE* **7** 1–10.

BOWDEN, J., SMITH, G. D. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525.

BOWDEN, J., SMITH, G. D., HAYCOCK, P. C. and BURGESS, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40** 304–314. https://doi.org/10.1002/gepi.21965

BOWMAN, L., HOPEWELL, J. C., CHEN, F., WALLENDSZUS, K., STEVENS, W., COLLINS, R. et al. and HPS3 AND TIMI55 REVEAL COLLABORATIVE GROUP (2017). Effects of anacetrapib in patients with atherosclerotic vascular disease. *N. Engl. J. Med.* **377** 1217–1227.

BURGESS, S., SCOTT, R. A., TIMPSON, N. J., SMITH, G. D., THOMPSON, S. G. and EPIC—INTERACT CONSORTIUM (2015). Using published data in Mendelian randomization: A blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30** 543–552. https://doi.org/10.1007/s10654-015-0011-z

Key words and phrases. Mendelian randomization, focused information criterion, postselection inference.

CASELLA, G. and HWANG, J. T. G. (2012). Shrinkage confidence procedures. *Statist. Sci.* **27** 51–60. MR2953495 https://doi.org/10.1214/10-STS319

DAVEY SMITH, G. and EBRAHIM, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32** 1–22.

DAVIES, N. M., VON HINKE KESSLER SCHOLDER, S., FARBMACHER, H., BURGESS, S., WINDMEIJER, F. and DAVEY SMITH, G. (2015). The many weak instruments problem and Mendelian randomization. *Stat. Med.* **34** 454–468. MR3301588 https://doi.org/10.1002/sim.6358

DITRAGLIA, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *J. Econometrics* **195** 187–208. MR3557268 https://doi.org/10.1016/j.jeconom.2016.07.006

DOBNIG, H., PILZ, S., SCHARNAGL, H., RENNER, W. and SEELHORST, U. WELLNITZ, B. KINKELDEI, J. BOEHM, B. O. WEIHRAUCH, G. et al. (2008). Independent association of low serum 25-hydroxyvitamin d and 1, 25-dihydroxyvitamin d levels with all-cause and cardiovascular mortality. *JAMA* **168** 1340–1349.

GILL, D., GEORGAKIS, M. K., WALKER, V. M., SCHMIDT, A. F., GKATZIONIS, A., DAVIES, N. M. et al. (2021). Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Res.* **6** 1–19.

HEMANI, G., BOWDEN, J. and DAVEY SMITH, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27** R195–R208. https://doi.org/10.1093/hmg/ddy163

HEMANI, G. et al. (2018). The MR-base platform supports systematic causal inference across the human phenome. *eLife* **7** 1–29.

JIANG, X., GE, T. and CHEN, C. Y. (2021). The causal role of circulating vitamin D concentrations in human complex traits and diseases: A large-scale Mendelian randomization study. *Sci. Rep.* **11** 1–10.

LASSI, G., TAYLOR, A. E., TIMPSON, N. J., KENNY, P. J., MATHER, R. J., EISEN, T. and MUNAFÒ, M. R. (2016). The CHRNA5-A3-B4 gene cluster and smoking: From discovery to therapeutics. *Trends Neurosci.* **39** 851–861.

LAWLOR, D. A., HARBORD, R. M., STERNE, J. A. C., TIMPSON, N. and SMITH, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27** 1133–1163. MR2420151 https://doi.org/10.1002/sim.3034

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

LEEB, H. and PÖTSCHER, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142** 201–211. MR2394290 https://doi.org/10.1016/j.jeconom.2007.05.017

MILLWOOD, I. Y., WALTERS, R. G., MEI, X. W., GUO, Y., YANG, L., BIAN, Z., BENNETT, D. A., CHEN, Y., DONG, C. et al. (2019). Conventional and genetic evidence on alcohol and vascular disease aetiology: A prospective study of 500,000 men and women in China. *Lancet* **393** 1831–1842.

MOKRY, L. E., ROSS, S., AHMAD, O. S., FORGETTA, V., SMITH, G. D., LEONG, A., GREENWOOD, C. M. T., THANASSOULIS, G. and RICHARDS, J. B. (2015). Vitamin D and risk of multiple sclerosis: A Mendelian randomization study. *PLoS Med.* **12** 1–20.

NEWEY, W. K. and WINDMEIJER, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica* **77** 687–719. MR2531359 https://doi.org/10.3982/ECTA6224

PATEL, A., DITRAGLIA, F. J., ZUBER, V. and BURGESS, S. (2024). Supplement to "Selecting invalid instruments to improve Mendelian randomization with two-sample summary data." https://doi.org/10.1214/23-AOAS1856SUPPA, https://doi.org/10.1214/23-AOAS1856SUPPB

REVEZ, J. A., LIN, T., QIAO, Z., XUE, A., HOLTZ, Y., ZHU, Z., ZENG, J., WANG, H., SIDORENKO, J. et al. (2020). Genome-wide association study identifies 143 loci associated with 25 hydroxyvitamin D concentration. *Nat. Commun.* **11** 1–12.

ROSENMAN, E. T. R., BASSE, G., OWEN, A. B. and BAIOCCHI, M. (2023). Combining observational and experimental datasets using shrinkage estimators. *Biometrics* **79** 2961–2973. MR4680697 https://doi.org/10.1111/biom.13827

SANDERSON, E., GLYMOUR, M. M., HOLMES, M. V., KANG, H., MORRISON, J., DAVEY SMITH, G. et al. (2022). Mendelian randomization. *Nat. Rev. Methods Primers* **2** 1–6.

SCHMIDT, A. F., FINAN, C. and GORDILLO-MARANON, M. ASSELBERGS, F. W. FREITAG, D. F. PATEL, R. S. et al. (2020). Genetic drug target validation using Mendelian randomization. *Nat. Commun.* **11** 1–12.

SCHMIDT, A. F., HUNT, N. B., GORDILLO-MARANON, M., CHAROEN, P., DRENOS, F., FINAN, C. et al. (2021). Cholesteryl Ester Transfer Protein (CETP) as a drug target for cardiovascular disease. *Nat. Commun.* **12** 1–10.

SOLOVIEFF, N., COTSAPAS, C., LEE, P. H., PURCELL, S. M. and SMOLLER, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.* **14** 483–495. https://doi.org/10.1038/nrg3461

STELZER, G. et al. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54** 1–30.

SWERDLOW, D. I., KUCHENBAECKER, K. B., SHAH, S., SOFAT, R., HOLMES, M. V., HINGORANI, A. D. et al. (2016). Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45** 1600–1616.

VERBANCK, M., CHEN, C.-Y., NEALE, B. and DO, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50** 693–698. https://doi.org/10.1038/s41588-018-0099-7

WILLIAMS, D. M., FINAN, C., SCHMIDT, A. F., BURGESS, S. and HINGORANI, A. D. (2020). Lipid lowering and Alzheimer's disease risk: A Mendelian randomization study. *Ann. Neurol.* **87** 30–39.

YE, T., SHAO, J. and KANG, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *Ann. Statist.* **49** 2079–2100. MR4319242 https://doi.org/10.1214/20-aos2027

ZHAO, Q., WANG, J., HEMANI, G., BOWDEN, J. and SMALL, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Ann. Statist.* **48** 1742–1769. MR4124342 https://doi.org/10.1214/19-AOS1866

The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1750–1772
https://doi.org/10.1214/23-AOAS1857

INVESTIGATING SWIMMING TECHNICAL SKILLS BY A DOUBLE PARTITION CLUSTERING OF MULTIVARIATE FUNCTIONAL DATA ALLOWING FOR DIMENSION SELECTION

BY ANTOINE BOUVET[1,a], SALIMA EL KOLEI[2,b] AND MATTHIEU MARBAC[2,c]

[1]*Université de Rennes, ENS Rennes, M2S Laboratory-EA 7470,* [a]*antoine.bouvet@ens-rennes.fr*
[2]*Université de Rennes Rennes, Ensai, CNRS, CREST—UMR 9194,* [b]*salima.el-kolei@ensai.fr,*
[c]*matthieu.marbac-lourdelle@ensai.fr*

Investigating technical skills of swimmers is a challenge for performance improvement that can be achieved by analyzing multivariate functional data recorded by inertial measurement units (IMU). To investigate technical levels of front-crawl swimmers, a new model-based approach is introduced to obtain two complementary partitions reflecting, for each swimmer, its swimming pattern and its ability to reproduce it. Contrary to the usual approaches for functional data clustering, the proposed approach also considers the information of the error terms resulting from the functional basis decomposition. Indeed, after decomposing into functional basis with finite number of elements both the original signal (measuring the swimming pattern) and the signal of squared error terms (measuring the ability to reproduce the swimming pattern), the method fits the joint distribution of the coefficients related to both decompositions by considering dependency between both partitions. Modeling this dependency is mandatory since the difficulty of reproducing a swimming pattern depends on its shape. Moreover, a sparse decomposition of the distribution within components that permits a selection of the relevant dimensions during clustering is proposed. The partitions obtained on the IMU data aggregate the kinematical stroke variability linked to swimming technical skills and allow relevant biomechanical strategy for front-crawl sprint performance to be identified.

REFERENCES

ABRAHAM, C., CORNILLON, P. A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30** 581–595. MR2002229 https://doi.org/10.1111/1467-9469.00350

BAUDRY, J.-P., RAFTERY, A. E., CELEUX, G., LO, K. and GOTTARDO, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.* **19** 332–353. MR2758307 https://doi.org/10.1198/jcgs.2010.08111

BOMPA, T. O. and BUZZICHELLI, C. (2018). *Periodization-: Theory and Methodology of Training*. Human Kinetics, Champaign.

BOUVET, A., EL KOLEI, S. and MARBAC, M. (2024). Supplement to "Investigating swimming technical skills by a double partition clustering of multivariate functional data allowing for dimension selection." https://doi.org/10.1214/23-AOAS1857SUPPA, https://doi.org/10.1214/23-AOAS1857SUPPB

BOUVEYRON, C., CELEUX, G., MURPHY, T. B. and RAFTERY, A. E. (2019). *Model-based Clustering and Classification for Data Science: With applications in R. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. MR3967046 https://doi.org/10.1017/9781108644181

BOUVEYRON, C., CÔME, E. and JACQUES, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *Ann. Appl. Stat.* **9** 1726–1760. MR3456352 https://doi.org/10.1214/15-AOAS861

BOUVEYRON, C. and JACQUES, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.* **5** 281–300. MR2860102 https://doi.org/10.1007/s11634-011-0095-6

BOUVEYRON, C., JACQUES, J., SCHMUTZ, A., SIMÕES, F. and BOTTINI, S. (2022). Co-clustering of multivariate functional data for the analysis of air pollution in the south of France. *Ann. Appl. Stat.* **16** 1400–1422. MR4455886 https://doi.org/10.1214/21-aoas1547

Key words and phrases. Clustering, feature selection, functional data, mixture models, sport performance.

CAMOMILLA, V., BERGAMINI, E., FANTOZZI, S. and VANNOZZI, G. (2018). Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: A systematic review. *Sensors* **18** 873. https://doi.org/10.3390/s18030873

DADASHI, F., MILLET, G. P. and AMINIAN, K. (2016). Front-crawl stroke descriptors variability assessment for skill characterisation. *J. Sports Sci.* **34** 1405–1412. https://doi.org/10.1080/02640414.2015.1114134

DELHAYE, E., BOUVET, A., NICOLAS, G., VILAS-BOAS, J. P., BIDEAU, B. and BIDEAU, N. (2022). Automatic Swimming Activity Recognition and Lap Time Assessment Based on a Single IMU: A Deep Learning Approach. *Sensors* **22**. https://doi.org/10.3390/s22155786

FERNANDES, A., GOETHEL, M., MARINHO, D. A., MEZÊNCIO, B., VILAS-BOAS, J. P. and FERNANDES, R. J. (2022a). Velocity Variability and Performance in Backstroke in Elite and Good-Level Swimmers. *Int. J. Environ. Res. Public Health* **19** 6744. https://doi.org/10.3390/ijerph19116744

FERNANDES, A., MEZÊNCIO, B., SOARES, S., DUARTE CARVALHO, D., SILVA, A., VILAS-BOAS, J. P. and FERNANDES, R. J. (2022b). Intra-and inter-cycle velocity variations in sprint front crawl swimming. *Sports Biomech.* 1–14. https://doi.org/10.1080/14763141.2022.2077815

FIGUEIREDO, P., KJENDLIE, P. L., VILAS-BOAS, J. P. and FERNANDES, R. J. (2012). Intracycle velocity variation of the body centre of mass in front crawl. *Int. J. Sports Med.* **33** 285–290. https://doi.org/10.1055/s-0031-1301323

FIGUEIREDO, P., PENDERGAST, D. R., VILAS-BOAS, J. P. and FERNANDES, R. J. (2013). Interplay of biomechanical, energetic, coordinative, and muscular factors in a 200 m front crawl swim. *BioMed Res. Int.* **2013**. https://doi.org/10.1155/2013/897232

FORRESTER, S. E. and TOWNEND, J. (2015). The effect of running velocity on footstrike angle–a curve-clustering approach. *Gait Posture* **41** 26–32. https://doi.org/10.1016/j.gaitpost.2014.08.004

GALIMBERTI, G., MANISI, A. and SOFFRITTI, G. (2018). Modelling the role of variables in model-based cluster analysis. *Stat. Comput.* **28** 145–169. MR3741643 https://doi.org/10.1007/s11222-017-9723-0

GALIMBERTI, G. and SOFFRITTI, G. (2007). Model-based methods to identify multiple cluster structures in a data set. *Comput. Statist. Data Anal.* **52** 520–536. MR2409999 https://doi.org/10.1016/j.csda.2007.02.019

GANZEVLES, S. P., BEEK, P. J., DAANEN, H. A., COOLEN, B. M. and TRUIJENS, M. J. (2019). Differences in swimming smoothness between elite and non-elite swimmers. *Sports Biomech.* 1–14. https://doi.org/10.1080/14763141.2019.1650102

GUIGNARD, B., ROUARD, A., CHOLLET, D. and SEIFERT, L. (2017). Behavioral dynamics in swimming: The appropriate use of inertial measurement units. *Front. Psychol.* **8** 383. https://doi.org/10.3389/fpsyg.2017.00383

HAMIDI RAD, M., GREMEAUX, V., DADASHI, F. and AMINIAN, K. (2020). A Novel Macro-micro Approach For Swimming Analysis In Main Swimming Techniques Using IMU Sensors. *Front. Bioeng. Biotechnol.* **8** 1511. https://doi.org/10.3389/fbioe.2020.597738

HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc.* **113** 649–659. MR3832216 https://doi.org/10.1080/01621459.2016.1273115

HELWIG, N. E., SHORTER, K. A., MA, P. and HSIAO-WECKSLER, E. T. (2016). Smoothing spline analysis of variance models: A new tool for the analysis of cyclic biomechanical data. *J. Biomech.* **49** 3216–3222. https://doi.org/10.1016/j.jbiomech.2016.07.035

HENNIG, C. (2010). Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.* **4** 3–34. MR2639661 https://doi.org/10.1007/s11634-010-0058-3

HENNIG, C. (2015). What are the true clusters? *Pattern Recogn. Lett.* **64** 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

IEVA, F., PAGANONI, A. M., PIGOLI, D. and VITELLI, V. (2011). Multivariate functional clustering for the analysis of ECG curves morphology. In *Cladag* 2011 (*8th International Meeting of the Classification and Data Analysis Group*) 1–4.

JACQUES, J. and PREDA, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing* **112** 164–171. https://doi.org/10.1016/j.neucom.2012.11.042

JACQUES, J. and PREDA, C. (2014a). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y

JACQUES, J. and PREDA, C. (2014b). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

LEROY, A. (2020). Multi-task learning models for functional data and application to the prediction of sports performances Ph.D. thesis Université de Paris.

LEROY, A., MARC, A., DUPAS, O., REY, J. L. and GEY, S. (2018). Functional Data Analysis in Sport Science: Example of Swimmers' Progression Curves Clustering. *Appl. Sci.* **8**. https://doi.org/10.3390/app8101766

LIEBL, D., WILLWACHER, S., HAMILL, J. and BRÜGGEMANN, G.-P. (2014). Ankle plantarflexion strength in rearfoot and forefoot runners: A novel clusteranalytic approach. *Hum. Mov. Sci.* **35** 104–120. https://doi.org/10.1016/j.humov.2014.03.008

MAGLISCHO, E. W. (2003). *Swimming Fastest*. Human Kinetics, Champaign.

MALLOR, F., LEON, T., GASTON, M. and IZQUIERDO, M. (2010). Changes in power curve shapes as an indicator of fatigue during dynamic contractions. *J. Biomech.* **43** 1627–1631. https://doi.org/10.1016/j.jbiomech.2010.01.038

MARBAC, M. and SEDKI, M. (2017). Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat. Comput.* **27** 1049–1063. MR3627562 https://doi.org/10.1007/s11222-016-9670-1

MARBAC, M., SEDKI, M. and PATIN, T. (2020). Variable selection for mixed data clustering: Application in human population genomics. *J. Classification* **37** 124–142. MR4111887 https://doi.org/10.1007/s00357-018-9301-y

MARBAC, M. and VANDEWALLE, V. (2019). A tractable multi-partitions clustering. *Comput. Statist. Data Anal.* **132** 167–179. MR3913142 https://doi.org/10.1016/j.csda.2018.06.013

MATSUDA, Y., YAMADA, Y., IKUTA, Y., NOMURA, T. and ODA, S. (2014). Intracyclic velocity variation and arm coordination for different skilled swimmers in the front crawl. *J. Human Kinet.* **44** 67. https://doi.org/10.2478/hukin-2014-0111

MOONEY, R., CORLEY, G., GODFREY, A., QUINLAN, L. R. and ÓLAIGHIN, G. (2015). Inertial Sensor Technology for Elite Swimming Performance Analysis: A Systematic Review. *Sensors* **16**. https://doi.org/10.3390/s16010018

PREATONI, E., HAMILL, J., HARRISON, A. J., HAYES, K., VAN EMMERIK, R. E., WILSON, C. and RODANO, R. (2013). Movement variability and skills monitoring in sports. *Sports Biomech.* **12** 69–92. https://doi.org/10.1080/14763141.2012.738700

RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. MR2268036 https://doi.org/10.1198/016214506000000113

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

RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 305–332. MR2188987 https://doi.org/10.1111/j.1467-9868.2006.00545.x

RIBEIRO, J., DE JESUS, K., FIGUEIREDO, P., TOUSSAINT, H., GUIDETTI, L., ALVES, F., VILAS-BOAS, J. P. and FERNANDES, R. (2013). Biomechanical determinants of force production in front crawl swimming. *J. Sports Med. Phys. Fit.* **53** 30–37.

SCHMUTZ, A., JACQUES, J., BOUVEYRON, C., CHEZE, L. and MARTIN, P. (2018). Données fonctionnelles multivariées issues d'objets connectés: Une méthode pour classer les individus. In *Journées des Statistiques*.

SCHMUTZ, A., JACQUES, J., BOUVEYRON, C., CHÉZE, L. and MARTIN, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Comput. Statist.* **35** 1101–1131. MR4133110 https://doi.org/10.1007/s00180-020-00958-4

SEIFERT, L., JESUS, K. D., KOMAR, J., RIBEIRO, J., ABRALDES, J. A., FIGUEIREDO, P., VILAS-BOAS, J. P. and FERNANDES, R. J. (2016). Behavioural variability and motor performance: Effect of practice specialization in front crawl swimming. *Hum. Mov. Sci.* **47** 141–150. https://doi.org/10.1016/j.humov.2016.03.007

SILVA, A. S., SALAZAR, A. J., BORGES, C. M. and CORREIA, M. V. (2011). Wearable monitoring unit for swimming performance analysis. In *International Joint Conference on Biomedical Engineering Systems and Technologies* 80–93. Springer, Berlin.

SLIMEN, Y. B., ALLIO, S. and JACQUES, J. (2018). Model-based co-clustering for functional data. *Neurocomputing* **291** 97–108. https://doi.org/10.1016/j.neucom.2018.02.055

STANIAK, Z., BUŚKO, K., GÓRSKI, M. and PASTUSZAK, A. (2016). Accelerometer profile of motion of the pelvic girdle in breaststroke swimming. *J. Human Kinet.* **52** 147. https://doi.org/10.1515/hukin-2016-0002

STANIAK, Z., BUŚKO, K., GÓRSKI, M. and PASTUSZAK, A. (2018). Accelerometer profile of motion of the pelvic girdle in butterfly swimming. *Acta Bioeng. Biomech.* **20** 159–167.

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563 https://doi.org/10.1198/016214504000001565

YAMAMOTO, M. (2012). Clustering of functional data in a low-dimensional subspace. *Adv. Data Anal. Classif.* **6** 219–247. MR2980648 https://doi.org/10.1007/s11634-012-0113-3

YAMAMOTO, M. and HWANG, H. (2017). Dimension-reduced clustering of functional data via subspace separation. *J. Classification* **34** 294–326. MR3669139 https://doi.org/10.1007/s00357-017-9232-z

YAMAMOTO, M. and TERADA, Y. (2014). Functional factorial K-means analysis. *Comput. Statist. Data Anal.* **79** 133–148. MR3227992 https://doi.org/10.1016/j.csda.2014.05.010