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*

Correction

THE ANNALS OF APPLIED STATISTICS Vol. 16, No. 1, pp. 1–625 March 2022

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

IMS PUBLICATIONS

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 16, Number 1, March 2022. 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
2022, Vol. 16, No. 1, 1–20
https://doi.org/10.1214/21-AOAS1493

ROBUST CAUSAL INFERENCE FOR INCREMENTAL RETURN ON AD SPEND WITH RANDOMIZED PAIRED GEO EXPERIMENTS

BY AIYOU CHEN[a] AND TIMOTHY C. AU[b]

Google LLC, [a]*aiyouchen@google.com,* [b]*timau@google.com*

Evaluating the incremental return on ad spend (iROAS) of a prospective online marketing strategy (i.e., the ratio of the strategy's causal effect on some response metric of interest relative to its causal effect on the ad spend) has become increasingly more important. Although randomized "geo experiments" are frequently employed for this evaluation, obtaining reliable estimates of iROAS can be challenging, as oftentimes only a small number of highly heterogeneous units are used. Moreover, advertisers frequently impose budget constraints on their ad spends which further complicates causal inference by introducing interference between the experimental units. In this paper we formulate a novel statistical framework for inferring the iROAS of online advertising from randomized paired geo experiment, which further motivates and provides new insights into Rosenbaum's arguments on instrumental variables, and we propose and develop a robust, distribution-free and interpretable estimator "Trimmed Match" as well as a data-driven choice of the tuning parameter which may be of independent interest. We investigate the sensitivity of Trimmed Match to some violations of its assumptions and show that it can be more efficient than some alternative estimators based on simulated data. We then demonstrate its practical utility with real case studies.

REFERENCES

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.

ATHEY, S., ECKLES, D. and IMBENS, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc.* **113** 230–240. MR3803460 https://doi.org/10.1080/01621459.2016.1241178

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550 https://doi.org/10.1198/jasa.2010.ap09490

BANG, H. and ZHAO, H. (2014). Cost-effectiveness analysis: A proposal of new reporting standards in statistical analysis. *J. Biopharm. Statist.* **24** 443–460. MR3196151 https://doi.org/10.1080/10543406.2013.860157

BLAIR, M. H. and KUSE, A. R. (2004). Better practices in advertising can change a cost of doing business to wise investments in the business. *J. Advert. Res.* **44** 71–89.

BLAKE, T., NOSKO, C. and TADELIS, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* **83** 155–174. MR3320110 https://doi.org/10.3982/ECTA12423

BLONIARZ, A., LIU, H., ZHANG, C.-H., SEKHON, J. S. and YU, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 7383–7390. MR3531136 https://doi.org/10.1073/pnas.1510506113

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. MR3341115 https://doi.org/10.1214/14-AOAS788

CHAUDHARY, M. A. and STEARNS, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: An example from a randomized trial. *Stat. Med.* **15** 1447–1458.

CHEN, A., LONGFILS, M. and BEST, C. (2020). The Python library for trimmed match and trimmed match design. Available at https://github.com/google/trimmed_match. [Online; accessed 2-June-2021].

CHEN, A., CHAN, D., PERRY, M., JIN, Y., SUN, Y., WANG, Y. and KOEHLER, J. (2018). Bias correction for paid search in media mix modeling. Technical Report, Google Inc. Available at https://research.google/pubs/pub46861.pdf.

Key words and phrases. Effect ratio, interference, heterogeneity, studentized trimmed mean.

DHAR, S. S. and CHAUDHURI, P. (2012). On the derivatives of the trimmed mean. *Statist. Sinica* **22** 655–679. MR2954356 https://doi.org/10.5705/ss.2010.155
DING, P., FELLER, A. and MIRATRIX, L. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797 https://doi.org/10.1111/rssb.12124
GOLDFARB, A. and TUCKER, C. (2011). Online advertising. In *Advances in Computers* (M. V. Zelkowitz, ed.) **81** 289–315. Elsevier. Chapter 6.
GORDON, B., ZETTELMEYER, F., BHARGAVA, N. and CHAPSKY, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Mark. Sci.* **38** 193–225.
HALL, P. (1981). Large sample properties of Jaeckel's adaptive trimmed mean. *Ann. Inst. Statist. Math.* **33** 449–462. MR0637757 https://doi.org/10.1007/BF02480955
HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2488795 https://doi.org/10.1002/9780470434697
HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292
IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751
INTERACTIVE ADVERTISING BUREAU (2018). IAB Internet advertising report: 2017 full year results.
JAECKEL, L. A. (1971). Some flexible estimates of location. *Ann. Math. Stat.* **42** 1540–1552. MR0350951 https://doi.org/10.1214/aoms/1177693152
JOHNSON, G. A., LEWIS, R. A. and NUBBEMEYER, E. I. (2017). Ghost ads: Improving the economics of measuring online ad effectiveness. *J. Mark. Res.* **54** 867–884.
KALYANAM, K., MCATEER, J., MAREK, J., HODGES, J. and LIN, L. (2018). Cross channel effects of search engine advertising on brick & mortar retail sales: Meta analysis of large scale field experiments on Google.com. *Quant. Mark. Econ.* **16** 1–42.
KERMAN, J., WANG, P. and VAVER, J. (2017). Estimating ad effectiveness using geo experiments in a time-based regression framework. Technical Report, Google, Inc. Available at https://research.google/pubs/pub45950.pdf.
KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562. MR0913573 https://doi.org/10.1214/aos/1176350609
KUENZEL, S. R. (2019). *Heterogeneous Treatment Effect Estimation Using Machine Learning*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Univ. California, Berkeley. MR4051203
LEHMANN, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*, 1st ed. Springer, New York. With the special assistance of H. J. M. D'Abrera. MR2279708
LEWIS, R. A. and RAO, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *Q. J. Econ.* **130** 1941–1973.
LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583
LUO, X., SMALL, D. S., LI, C.-S. R. and ROSENBAUM, P. R. (2012). Inference with interference between units in an fMRI experiment of motor inhibition. *J. Amer. Statist. Assoc.* **107** 530–541. MR2980065 https://doi.org/10.1080/01621459.2012.655954
NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076
ROLNICK, D., AYDIN, K., POUGET-ABADIE, J., KAMALI, S., MIRROKNI, V. and NAJMI, A. (2019). Randomized experimental design via geographic clustering. In *Proceedings of the 25th ACM International Conference on Knowledge Discovery & Data Mining* 2745–2753. https://doi.org/10.1145/3292500.3330778
ROSENBAUM, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *J. Amer. Statist. Assoc.* **91** 444–444.
ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. With comments and a rejoinder by the author. MR1962487 https://doi.org/10.1214/ss/1042727942
ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112
ROSENBAUM, P. R. (2020). *Design of Observational Studies. Springer Series in Statistics*. Springer, Cham. Second edition [of 2561612]. MR4225301 https://doi.org/10.1007/978-3-030-46405-9
RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *J. Amer. Statist. Assoc.* **75** 591–593.
SAPP, S., VAVER, J., SCHURINGA, J. and DROPSHO, S. (2017). Near impressions for observational causal ad impact. Technical Report, Google Inc. Available at https://research.google/pubs/pub46418.pdf.
SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *J. Amer. Statist. Assoc.* **103** 924–933. MR2528819 https://doi.org/10.1198/016214507000001247

TUKEY, J. W. and MCLAUGHLIN, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. I. *Sankhyā Ser. A* **25** 331–352. MR0169354

VARIAN, H. R. (2009). Online ad auctions. *Am. Econ. Rev.* **99** 430–34. https://doi.org/10.1257/aer.99.2.430

VARIAN, H. R. (2016). Causal inference in economics and marketing. *Proc. Natl. Acad. Sci. USA* **113** 7310–7315.

VAVER, J. and KOEHLER, J. (2011). Measuring ad effectiveness using geo experiments. Technical Report, Google Inc. Available at https://research.google/pubs/pub38355.pdf.

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

WICKHAM, H. (2016). *Ggplot2*: *Elegant Graphics for Data Analysis*. Springer, New York.

YE, Q., MALIK, S., CHEN, J. and ZHU, H. (2016). The seasonality of paid search effectiveness from a long running field test. In *Proceedings of the* 2016 *ACM Conference on Economics and Computation*. *EC* '16 515–530. ACM, New York.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 21–39
https://doi.org/10.1214/21-AOAS1492

BAGEL: A BAYESIAN GRAPHICAL MODEL FOR INFERRING DRUG EFFECT LONGITUDINALLY ON DEPRESSION IN PEOPLE WITH HIV

BY YULIANG LI[1,a], YANG NI[2,c], LEAH H. RUBIN[3,d], AMANDA B. SPENCE[4,e] AND YANXUN XU[1,b]

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University,* [a]*yli193@jhu.edu,* [b]*yanxun.xu@jhu.edu*
[2]*Department of Statistics, Texas A&M University,* [c]*yni@stat.tamu.edu*
[3]*Departments of Neurology and Psychiatry, Johns Hopkins University School of Medicine,* [d]*lrubin@jhu.edu*
[4]*Department of Medicine, Georgetown University,* [e]*abs132@georgetown.edu*

Access and adherence to antiretroviral therapy (ART) has transformed the face of HIV infection from a fatal to a chronic disease. However, ART is also known for its side effects. Studies have reported that ART is associated with depressive symptomatology. Large-scale HIV clinical databases with individuals' longitudinal depression records, ART medications, and clinical characteristics offer researchers unprecedented opportunities to study the effects of ART drugs on depression over time. We develop BAGEL, a Bayesian graphical model, to investigate longitudinal effects of ART drugs on a range of depressive symptoms while adjusting for participants' demographic, behavior, and clinical characteristics, and taking into account the heterogeneous population through a Bayesian nonparametric prior. We evaluate BAGEL through simulation studies. Application to a dataset from the Women's Interagency HIV Study yields interpretable and clinically useful results. BAGEL not only can improve our understanding of ART drugs' effects on disparate depression symptoms but also has clinical utility in guiding informed and effective treatment selection to facilitate precision medicine in HIV.

REFERENCES

ABERS, M. S., SHANDERA, W. X. and KASS, J. S. (2014). Neurological and psychiatric adverse effects of antiretroviral drugs. *CNS Drugs* **28** 131–145.

ADIMORA, A. A., RAMIREZ, C., BENNING, L., GREENBLATT, R. M., KEMPF, M.-C., TIEN, P. C., KASSAYE, S. G., ANASTOS, K., COHEN, M. et al. (2018). Cohort profile: The women's interagency HIV study (WIHS). *Int. J. Epidemiol.* **47** 393–394. https://doi.org/10.1093/ije/dyy021

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

ARENAS-PINTO, A., GRUND, B., SHARMA, S., MARTINEZ, E., CUMMINS, N., FOX, J., KLINGMAN, K. L., SEDLACEK, D., COLLINS, S. et al. (2018). Risk of suicidal behavior with use of efavirenz: Results from the strategic timing of antiretroviral treatment trial. *Clin. Infect. Dis.* **67** 420–429. https://doi.org/10.1093/cid/ciy051

BACON, M. C., VON WYL, V., ALDEN, C., SHARP, G., ROBISON, E., HESSOL, N., GANGE, S., BARRANDAY, Y., HOLMAN, S. et al. (2005). The women's interagency hiv study: An observational cohort brings clinical sciences to the bench. *Clin. Diagn. Lab. Immunol.* **12** 1013–1019.

BARKAN, S. E., MELNICK, S. L., PRESTON-MARTIN, S., WEBER, K., KALISH, L. A., MIOTTI, P., YOUNG, M., GREENBLATT, R., SACKS, H. et al. (1998). The women's interagency hiv study. *Epidemiology* 117–125.

BENGTSON, A. M., PENCE, B. W., CRANE, H. M., CHRISTOPOULOS, K., FREDERICKSEN, R. J., GAYNES, B. N., HEINE, A., MATHEWS, W. C., MOORE, R. et al. (2016). Disparities in depressive symptoms and antidepressant treatment by gender and race/ethnicity among people living with HIV in the United States. *PLoS ONE* **11** e0160738. https://doi.org/10.1371/journal.pone.0160738

Key words and phrases. Bayesian nonparametrics, depression, graphical model, longitudinal cohort study, precision medicine.

BENGTSON, A. M., PENCE, B. W., MOLLAN, K. R., EDWARDS, J. K., MOORE, R. D. and O'CLEIRIGH, C. (2017). The relationship between efavirenz as initial antiretroviral therapy and suicidal thoughts among HIV-infected adults in routine care. *Journal of Acquired Immune Deficiency Syndromes* (1999) **76** 402.

BEST, B. M., LETENDRE, S. L., KOOPMANS, P., ROSSI, S. S., CLIFFORD, D. B., COLLIER, A. C., GELMAN, B. B., MARRA, C. M., MCARTHUR, J. C. et al. (2012). Low csf concentrations of the nucleotide hiv reverse transcriptase inhibitor, tenofovir. *Journal of Acquired Immune Deficiency Syndromes* (1999) **59** 376.

BORGHETTI, A., BALDIN, G., CAPETTI, A., STERRANTCOHENINO, G., RUSCONI, S., LATINI, A., GIACOMETTI, A., MADEDDU, G., PICARELLI, C. et al. (2017). Efficacy and tolerability of dolutegravir and two nucleos (t) ide reverse transcriptase inhibitors in HIV-1-positive, virologically suppressed patients. *AIDS* **31** 457–459.

BORGHETTI, A., CALCAGNO, A., LOMBARDI, F., CUSATO, J., BELMONTI, S., D?AVOLIO, A., CICCARELLI, N., LA MONICA, S., COLAFIGLI, M. et al. (2018). SLC22A2 variants and dolutegravir levels correlate with psychiatric symptoms in persons with HIV. *Journal of Antimicrobial Chemotherapy*.

BRICKMAN, C., PROPERT, K. J., VOYTEK, C., METZGER, D. and GROSS, R. (2017). Association between depression and condom use differs by sexual behavior group in patients with HIV. *AIDS Behav.* **21** 1676–1683. https://doi.org/10.1007/s10461-016-1610-8

BRINK, M. S., VISSCHER, C., ARENDS, S., ZWERVER, J., POST, W. J. and LEMMINK, K. A. (2010). Monitoring stress and recovery: New insights for the prevention of injuries and illnesses in elite youth soccer players. *British Journal of Sports Medicine* **44** 809–815.

CHATTOPADHYAY, S., BALL, S., KARGUPTA, A., TALUKDAR, P., ROY, K., TALUKDAR, A. and GUHA, P. (2017). Cognitive behavioral therapy improves adherence to antiretroviral therapy in hiv-infected patients: A prospective randomized controlled trial from eastern India. *HIV & AIDS Review. International Journal of HIV-Related Problems* **16** 89–95.

CLUBRETH, R., DUBE, S. and MAGGIO, D. (2016). Associations between major depression, health-risk behaviors, and medication adherence among hiv-positive adults receiving medical care in Georgia. *Journal of the Georgia Public Health Association*.

COHEN, C., ELION, R., RUANE, P., SHAMBLAW, D., DEJESUS, E., RASHBAUM, B., CHUCK, S. L., YALE, K., LIU, H. C. et al. (2011). Randomized, phase 2 evaluation of two single-tablet regimens elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate versus efavirenz/emtricitabine/tenofovir disoproxil fumarate for the initial treatment of hiv infection. *AIDS* **25** F7–F12.

COHEN, J., D?AGOSTINO, L., WILSON, J., TUZER, F. and TORRES, C. (2017). Astrocyte senescence and metabolic changes in response to hiv antiretroviral therapy drugs. *Front. Aging Neurosci.* **9** 281.

COOK, J. A., COHEN, M. H., BURKE, J., GREY, D., ANASTOS, K., KIRSTEIN, L., PALACIO, H., RICHARDSON, J., WILSON, T. et al. (2002). Effects of depressive symptoms and mental health quality of life on use of highly active antiretroviral therapy among hiv-seropositive women. *Jaids-Hagerstown MD* **30** 401–409.

DEROGATIS, L. R., LIPMAN, R. S., RICKELS, K., UHLENHUTH, E. H. and COVI, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral Science* **19** 1–15.

DEY, S., ZHANG, P., SOW, D. and NG, K. (2019). Perdrep: Personalized drug effectiveness prediction from longitudinal observational data. In *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 1258–1268.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. With comments and a rejoinder by the authors. MR1435485 https://doi.org/10.1214/ss/1038425655

ELZI, L., ERB, S., FURRER, H., CAVASSINI, M., CALMY, A., VERNAZZA, P., GÜNTHARD, H., BERNASCONI, E. and BATTEGAY, M. (2017). Adverse events of raltegravir and dolutegravir. *AIDS* (*London*, *England*) **31** 1853.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. MR0438568

FRIED, E. I., VAN BORKULO, C. D., EPSKAMP, S., SCHOEVERS, R. A., TUERLINCKX, F. and BORSBOOM, D. (2016). Measuring depression over time... Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychol. Assess* **28** 1354–1367. https://doi.org/10.1037/pas0000275

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

HARRIS, M., LARSEN, G. and MONTANER, J. S. (2008). Exacerbation of depression associated with starting raltegravir: A report of four cases. *AIDS* **22** 1890–1892.

HOFFMANN, C., WELZ, T., SABRANSKI, M., KOLB, M., WOLF, E., STELLBRINK, H.-J. and WYEN, C. (2017). Higher rates of neuropsychiatric adverse events leading to dolutegravir discontinuation in women and older patients. *HIV Med.* **18** 56–63. https://doi.org/10.1111/hiv.12468

ICKOVICS, J. R., HAMBURGER, M. E., VLAHOV, D., SCHOENBAUM, E. E., SCHUMAN, P., BOLAND, R. J., MOORE, J. and HIV EPIDEMIOLOGY RESEARCH STUDY GROUP (2001). Mortality, cd4 cell count decline, and depressive symptoms among hiv-seropositive women: Longitudinal analysis from the hiv epidemiology research study. *JAMA* **285** 1466–1474.

IRONSON, G., FITCH, C. and STUETZLE, R. (2017). Depression and survival in a 17-year longitudinal study of people with hiv: Moderating effects of race and education. *Psychosomatic Medicine* **79** 749–756.

JONES, C. M., GRIFFITHS, P. C. and MELLALIEU, S. D. (2017). Training load and fatigue marker associations with injury and illness: A systematic review of longitudinal studies. *Sports Med.* **47** 943–974. https://doi.org/10.1007/s40279-016-0619-5

KAPFHAMMER, H.-P. (2006). Somatic symptoms in depression. *Dialogues Clin. Neurosci.* **8** 227–239.

KROENKE, K., SPITZER, R. L. and WILLIAMS, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16** 606–613.

LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877 https://doi.org/10.1198/1061860043010

LEWINSOHN, P. M., SEELEY, J. R., ROBERTS, R. E. and ALLEN, N. B. (1997). Center for epidemiologic studies depression scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychol. Aging* **12** 277–287. https://doi.org/10.1037//0882-7974.12.2.277

LI, Y., BANDYOPADHYAY, D., XIE, F. and XU, Y. (2020). BAREB: A Bayesian repulsive biclustering model for periodontal data. *Stat. Med.* **39** 2139–2151. MR4108756 https://doi.org/10.1002/sim.8536

LI, Y., NI, Y., RUBIN, L. H., SPENCE, A. B. and XU, Y. (2022). Supplement 1 to "BAGEL: A Bayesian graphical model for inferring drug effect longitudinally on depression in people with HIV." https://doi.org/10.1214/21-AOAS1492SUPPA

LI, Y., NI, Y., RUBIN, L. H., SPENCE, A. B. and XU, Y. (2022). Supplement 2 to "BAGEL: A Bayesian graphical model for inferring drug effect longitudinally on depression in people with HIV." https://doi.org/10.1214/21-AOAS1492SUPPB

LIU, B., LI, Y., GHOSH, S., SUN, Z., NG, K. and HU, J. (2019). Complication risk profiling in diabetes care: A Bayesian multi-task and feature relationship learning approach. *IEEE Trans. Knowl. Data Eng.*

MAKI, P. M., RUBIN, L. H., COHEN, M., GOLUB, E. T., GREENBLATT, R. M., YOUNG, M., SCHWARTZ, R. M., ANASTOS, K. and COOK, J. A. (2012). Depressive symptoms are increased in the early perimenopausal stage in ethnically diverse HIV+ and HIV- women. *Menopause* (*New York, NY*) **19** 1215.

MILLS, A., ARRIBAS, J. R., ANDRADE-VILLANUEVA, J., DIPERRI, G., VAN LUNZEN, J., KOENIG, E., ELION, R., CAVASSINI, M., MADRUGA, J. V. et al. (2016). Switching from tenofovir disoproxil fumarate to tenofovir alafenamide in antiretroviral regimens for virologically suppressed adults with hiv-1 infection: A randomised, active-controlled, multicentre, open-label, phase 3, non-inferiority study. *Lancet Infect. Dis.* **16** 43–52.

MOLLAN, K. R., SMURZYNSKI, M., ERON, J. J., DAAR, E. S., CAMPBELL, T. B., SAX, P. E., GULICK, R. M., NA, L., O'KEEFE, L. et al. (2014). Association between efavirenz as initial therapy for hiv-1 infection and increased risk for suicidal ideation or attempted or completed suicide: An analysis of trial data. *Ann. Intern. Med.* **161** 1–10.

MOLLAN, K. R., TIERNEY, C., HELLWEGE, J. N., ERON, J. J., HUDGENS, M. G., GULICK, R. M., HAUBRICH, R., SAX, P. E., CAMPBELL, T. B. et al. (2017). Race/ethnicity and the pharmacogenetics of reported suicidality with efavirenz among clinical trials participants. *J. Infect. Dis.* **216** 554–564.

MOORE, J., SCHUMAN, P., SCHOENBAUM, E., BOLAND, B., SOLOMON, L. and SMITH, D. (1999). Severe adverse life events and depressive symptoms among women with, or at risk for, hiv infection in four cities in the United States of America. *AIDS* **13** 2459–2468.

MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95–110. MR2082149 https://doi.org/10.1214/088342304000000017

MÜLLER, P. and VANNUCCI, M., eds. (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge Univ. Press, Cambridge. MR2269095 https://doi.org/10.1017/CBO9780511584589

NANNI, M. G., CARUSO, R., MITCHELL, A. J., MEGGIOLARO, E. and GRASSI, L. (2015). Depression in HIV infected patients: A review. *Curr. Psychiatry Rep.* **17** 530. https://doi.org/10.1007/s11920-014-0530-4

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News* **6** 7–11.

REVUELTA-HERRERO, J. L., CHAMORRO-DE VEGA, E., RODRÍGUEZ-GONZÁLEZ, C. G., ALONSO, R., HERRANZ-ALONSO, A. and SANJURJO-SÁEZ, M. (2018). Effectiveness, safety, and costs of a treatment switch to dolutegravir plus rilpivirine dual therapy in treatment-experienced hiv patients. *Annals of Pharmacotherapy* **52** 11–18.

RUBIN, L. H., COOK, J. A., GREY, D. D., WEBER, K., WELLS, C., GOLUB, E. T., WRIGHT, R. L., SCHWARTZ, R. M., GOPARAJU, L. et al. (2011). Perinatal depressive symptoms in HIV-infected versus HIV-uninfected women: A prospective study from preconception to postpartum. *Journal of Women's Health* **20** 1287–1295.

SHAH, A., GANGWANI, M. R., CHAUDHARI, N. S., GLAZYRIN, A., BHAT, H. K. and KUMAR, A. (2016). Neurotoxicity in the post-haart era: Caution for the antiretroviral therapeutics. *Neurotoxicity Research* **30** 677–697.

SQUIRES, K., POZNIAK, A. L., PIERONE, G., STEINHART, C. R., BERGER, D., BELLOS, N. C., BECKER, S. L., WULFSOHN, M., MILLER, M. D. et al. (2003). Tenofovir disoproxil fumarate in nucleoside-resistant hiv-1 infection: A randomized trial. *Ann. Intern. Med.* **139** 313–320.

TAIBI, D. M. (2013). Sleep disturbances in persons living with hiv. *Journal of the Association of Nurses in AIDS Care* **24** S72–S85.

TANIGUCHI, T., SHACHAM, E., ÖNEN, N. F., GRUBB, J. R. and OVERTON, E. T. (2014). Depression severity is associated with increased risk behaviors and decreased cd4 cell counts. *AIDS Care* **26** 1004–1012.

UNDERWOOD, J., ROBERTSON, K. R. and WINSTON, A. (2015). Could antiretroviral neurotoxicity play a role in the pathogenesis of cognitive impairment in treated hiv disease? *AIDS* **29** 253–261.

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

WILLIAMS, D. W., LI, Y., DASTGHEYB, R., FITZGERALD, K. C., MAKI, P. M., SPENCE, A. B., GUSTAFSON, D. R., MILAM, J., SHARMA, A. et al. (2020). Associations between antiretroviral drugs on depressive symptomatology in homogenous subgroups of women with hiv. *Journal of Neuroimmune Pharmacology* 1–14.

XU, Y., XU, Y. and SARIA, S. (2016). A non-parametric Bayesian approach for estimating treatment-response curves from sparse time series. In *Proceedings of the* 1*st Machine Learning for Healthcare Conference* 282–300.

ZASH, R., MAKHEMA, J. and SHAPIRO, R. L. (2018). Neural-tube defects with dolutegravir treatment from the time of conception. *N. Engl. J. Med.* **379** 979–981.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 40–59
https://doi.org/10.1214/21-AOAS1468

SUBGROUP IDENTIFICATION AND VARIABLE SELECTION FOR TREATMENT DECISION MAKING

BY BAQUN ZHANG[1,a] AND MIN ZHANG[2,b]

[1]*School of Statistics and Management, Shanghai University of Finance and Economics,* [a]*zhang.baqun@mail.shufe.edu.cn*
[2]*Department of Biostatistics, University of Michigan,* [b]*mzhangst@umich.edu*

When treatment effect heterogeneity exists, identifying the subgroup of patients who would benefit from an active treatment relative to a control is an important question. This article focuses on subgroup identification in the presence of a large dimensional set of covariates, with the number of covariates possibly greater than the sample size. We approach this problem from the perspective of optimal treatment decision rules and propose methods that can simultaneously estimate the treatment decision rule and select prescriptive variables important for treatment decision making and subgroup identification. The proposed methods are built within a robust classification framework based on doubly robust augmented inverse probability weighted estimators (AIPWE), hence sharing the robustness property. An L_1 (lasso-type) penalty is used within the classification framework to target selection of prescriptive variables. We further propose a backward elimination process for fine-tuning selection. The methods can be conveniently implemented by taking advantage of standard software for logistic regression and lasso. The methods are evaluated by extensive simulation studies which demonstrated the superior and robust performance of the proposed methods relative to existing ones. In addition, the estimated decision rules from the proposed methods are considerably simpler than other methods. We applied various methods to identify the subgroup of patients suitable for each of the two commonly used anticoagulants in terms of bleeding risk for patients with acute myocardial infarction undergoing percutaneous coronary intervention.

REFERENCES

ANDREOU, C., MANIOTIS, C. and KOUTOUZIS, M. (2017). The rise and fall of anticoagulation with bivalirudin during percutaneous coronary interventions: A review article. *Cardiol Ther* **6** 1–12. https://doi.org/10.1007/s40119-017-0082-x

ATHEY, S. and IMBENS, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *Stat* **1050** 1–26.

ATHEY, S. and WAGER, S. (2021). Policy learning with observational data. *Econometrica* **89** 133–161. MR4220385 https://doi.org/10.3982/ecta15732

BARGAGLI STOFFI, F., TORTÚ, C. and FORASTIERE, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference. Available at arXiv:2008.00707.

BARRETT, J. K., HENDERSON, R. and ROSTHØJ, S. (2014). Doubly robust estimation of optimal dynamic treatment regimes. *Stat. Biosci.* **6** 244–260.

BIERNOT, P. and MOODIE, E. E. M. (2010). A comparison of variable selection approaches for dynamic treatment regimes. *Int. J. Biostat.* **6** Art. 6, 20. MR2594878 https://doi.org/10.2202/1557-4679.1178

BRINKLEY, J., TSIATIS, A. and ANSTROM, K. J. (2010). A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics* **66** 512–522. MR2758831 https://doi.org/10.1111/j.1541-0420.2009.01282.x

CAVENDER, M. A. and SABATINE, M. S. (2014). Bivalirudin versus heparin in patients planned for percutaneous coronary intervention: A meta-analysis of randomised controlled trials. *Lancet* **384** 599–606.

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

Key words and phrases. Backward elimination, classification, effect modification, heterogeneous treatment effect, lasso, interaction, optimal treatment regime, variable selection.

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534 https://doi.org/10.1111/biom.12676
ELGENDY, I. Y. and CAPODANNO, D. (2017). Heparin versus bivalirudin for percutaneous coronary intervention: Has the debate come to an end? *Journal of Thoracic Disease* **9** 4305–4307.
FAN, A., LU, W. and SONG, R. (2016). Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat.* **10** 32–53. MR3480486 https://doi.org/10.1214/15-AOAS849
FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689 https://doi.org/10.1002/sim.4322
GUNTER, L., CHERNICK, M. and SUN, J. (2011). A simple method for variable selection in regression with respect to treatment selection. *Pak. J. Stat. Oper. Res.* 7(2–Sp).
GUNTER, L., ZHU, J. and MURPHY, S. A. (2011). Variable selection for qualitative interactions. *Stat. Methodol.* **8** 42–55. MR2741508 https://doi.org/10.1016/j.stamet.2009.05.003
KOSOROK, M. R. and MOODIE, E. E. M., eds. (2016). *Adaptive Treatment Strategies in Practice*: *Planning Trials and Analyzing Data for Personalized Medicine*. *ASA-SIAM Series on Statistics and Applied Probability*. SIAM, Philadelphia, PA. MR3450070
LU, W., ZHANG, H. H. and ZENG, D. (2013). Variable selection for optimal treatment decision. *Stat. Methods Med. Res.* **22** 493–504. MR3190671 https://doi.org/10.1177/0962280211428383
LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384
MOODIE, E. E. M., RICHARDSON, T. S. and STEPHENS, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455. MR2370803 https://doi.org/10.1111/j.1541-0420.2006.00686.x
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
PERDONCIN, E., ZHANG, M., RIBA, A., LALONDE, T. A., GRINES, C. L. and GURM, H. S. (2013). Impact of worsening renal dysfunction on the comparative efficacy of bivalirudin and platelet glycoprotein IIb/IIIa inhibitors: Insights from Blue Cross Blue Shield of Michigan Cardiovascular Consortium. *Circulation*: *Cardiovascular Interventions* **6** 688–693.
QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864
ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27** 4678–4721. MR2528576 https://doi.org/10.1002/sim.3301
SHI, C., LU, W. and SONG, R. (2020). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *J. Mach. Learn. Res.* **21** Paper No. 176, 67. MR4209462
SHI, C., SONG, R. and LU, W. (2019). On testing conditional qualitative treatment effects. *Ann. Statist.* **47** 2348–2377. MR3953454 https://doi.org/10.1214/18-AOS1750
SONG, R., KOSOROK, M., ZENG, D., ZHAO, Y., LABER, E. and YUAN, M. (2015). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat* **4** 59–68. MR3405390 https://doi.org/10.1002/sta4.78
TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. MR3293607 https://doi.org/10.1080/01621459.2014.951443
TSIATIS, A. A., DAVIDIAN, M., HOLLOWAY, S. T. and LABER, E. B. (2019). *Dynamic Treatment Regime*: *Statistical Methods for Precision Medicine*. Chapman & Hall.
WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839
WANG, L., ZHOU, Y., SONG, R. and SHERWOOD, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113** 1243–1254. MR3862354 https://doi.org/10.1080/01621459.2017.1330204
ZHANG, B. and ZHANG, M. (2018a). Variable selection for estimating the optimal treatment regimes in the presence of a large number of covariates. *Ann. Appl. Stat.* **12** 2335–2358. MR3875703 https://doi.org/10.1214/18-AOAS1154
ZHANG, B. and ZHANG, M. (2018b). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74** 891–899. MR3860710 https://doi.org/10.1111/biom.12836
ZHANG, B. and ZHANG, M. (2022). Supplement to "Subgroup identification and variable selection for treatment decision making." https://doi.org/10.1214/21-AOAS1468SUPPA, https://doi.org/10.1214/21-AOAS1468SUPPB
ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012a). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x
ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. (2012b). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114. MR4027418 https://doi.org/10.1002/sta.411

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 60–79
https://doi.org/10.1214/21-AOAS1476

BOUNDING THE LOCAL AVERAGE TREATMENT EFFECT IN AN INSTRUMENTAL VARIABLE ANALYSIS OF ENGAGEMENT WITH A MOBILE INTERVENTION

BY ANDREW J. SPIEKER[1,a], ROBERT A. GREEVY[1,b], LYNDSAY A. NELSON[2,c] AND LINDSAY S. MAYBERRY[2,d]

[1]*Department of Biostatistics, Vanderbilt University Medical Center,* [a]*andrew.spieker@vumc.org,* [b]*robert.greevy@vumc.org*
[2]*Department of Medicine, Vanderbilt University Medical Center,* [c]*lyndsay.a.nelson@vumc.org,* [d]*lindsay.mayberry@vumc.org*

Estimation of local average treatment effects in randomized trials typically relies upon the exclusion restriction assumption in cases where we are unwilling to rule out the possibility of unmeasured confounding. Under this assumption, treatment effects are mediated through the post-randomization variable being conditioned upon and directly attributable to neither the randomization itself nor its latent descendants. Recently, there has been interest in mobile health interventions to provide healthcare support. Mobile health interventions (e.g., the Rapid Encouragement/Education and Communications for Health, or REACH, designed to support self management for adults with type 2 diabetes) often involve both one-way and interactive messages. In practice, it is highly likely that any benefit from the intervention is achieved both through receipt of the intervention content and through engagement with/response to it. Application of an instrumental variable analysis in order to understand the role of engagement with REACH (or a similar intervention) requires the traditional exclusion restriction assumption to be relaxed. We propose a conceptually intuitive sensitivity analysis procedure for the REACH randomized trial that places bounds on local average treatment effects. Simulation studies reveal this approach to have desirable finite-sample behavior and to recover local average treatment effects under correct specification of sensitivity parameters.

REFERENCES

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

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

BUSE, A. (1992). The bias of instrumental variable estimators. *Econometrica* **60** 173–180. MR1161549 https://doi.org/10.2307/2951682

DORIE, V., HARADA, M., CARNEGIE, N. B. and HILL, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat. Med.* **35** 3453–3470. MR3537215 https://doi.org/10.1002/sim.6973

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* **1** 54–77. With a comment by J. A. Hartigan and a rejoinder by the authors. MR0833275

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., RUBIN, D. and ZHOU, X. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics* **3** 147–164.

FUNK, M., WESTREICH, D., WIESEN, C., STÜRMER, T., BROOKHART, M. and DAVIDIAN, M. (2011). Doubly robust estimation of causal effects. *Am. J. Epidemiol.* **173** 761–767.

Key words and phrases. Causal, exclusion restriction, instrumental variables, sensitivity analysis.

GREENWOOD, D. A., GEE, P. M., FATKIN, K. J. and PEEPLES, M. (2017). A systematic review of reviews evaluating technology-enabled diabetes self-management education and support. *J. Diabetes Sci. Technol.* **11** 1015–1027. https://doi.org/10.1177/1932296817713506

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

HARRELL, F. (2001). *Regression Modeling Strategies*: *With Applications to Linear Models*, *Logistic Regression*, *and Survival Analysis*. Springer, New York.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.* **25** 51–71. MR2741814 https://doi.org/10.1214/10-STS321

IMBENS, G. and ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

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

JO, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *J. Educ. Behav. Stat.* **27** 385–409.

JO, B. (2007). Bias mechanisms in intention-to-treat analysis with data subject to treatment noncompliance and missing outcomes. *J. Educ. Behav. Stat.* **33** 158–185. https://doi.org/10.3102/1076998607302635

JO, B. and VINOKUR, A. D. (2011). Sensitivity analysis and bounding of causal effects with alternative identifying assumptions. *J. Educ. Behav. Stat.* **36** 415–440. https://doi.org/10.3102/1076998610383985

LIN, D., PSATY, B. and KRONMAL, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

LUNCEFORD, J. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

MARCOLINO, M., OLIVEIRA, J., D'AGOSTINO, M., RIBEIRO, A., ALKMIM, M. and NOVILLO-ORTIZ, D. (2018). The impact of mHealth interventions: Systematic review of systematic reviews. *Journal of Medical Internet Research MHealth UHealth* **6** e23.

MILLIMET, D. L. and TCHERNIS, R. (2013). Estimation of treatment effects without an exclusion restriction: With an application to the analysis of the school breakfast program. *J. Appl. Econometrics* **28** 982–1017. MR3108036

NELSON, L., WALLSTON, K., KRIPALANI, S., GREEVY, R. J., ELASY, T., BERGNER, E., GENTRY, C. and MAYBERRY, L. (2018). Mobile phone support for diabetes self-care among diverse adults: Protocol for a three-arm randomized controlled trial. *JMIR Res. Protoc.* **7** e92.

NELSON, L. A., GREEVY, R. A., SPIEKER, A., WALLSTON, K. A., ELASY, T. A., KRIPALANI, S., GENTRY, C., BERGNER, E. M., LESTOURGEON, L. M. et al. (2021). Effects of a tailored text messaging intervention among diverse adults with type 2 diabetes: Evidence from the 15-month REACH randomized controlled trial. *Diabetes Care* **44** 26–34. https://doi.org/10.2337/dc20-0961

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

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. *Mathematical Modeling* **7** 1393–1512.

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

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROY, J., HOGAN, J. and MARCUS, B. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics* **9** 277–289.

RUBIN, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

SCHOMAKER, M. and HEUMANN, C. (2018). Bootstrap inference when using multiple imputation. *Stat. Med.* **37** 2252–2266. MR3810720 https://doi.org/10.1002/sim.7654

SPIEKER, A. J., GREEVY, R. A., NELSON, L. A. and MAYBERRY, L. S. (2022). Supplement to "Bounding the local average treatment effect in an instrumental variable analysis of engagement with a mobile intervention." https://doi.org/10.1214/21-AOAS1476SUPPA, https://doi.org/10.1214/21-AOAS1476SUPPB

STUART, E. A. and JO, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Stat. Methods Med. Res.* **24** 657–674. MR3428422 https://doi.org/10.1177/0962280211421840

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. MR2528575 https://doi.org/10.1002/sim.3113

VAN BUUREN, S., BRAND, J. P. L., GROOTHUIS-OUDSHOORN, C. G. M. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76** 1049–1064. MR2307507 https://doi.org/10.1080/10629360600810434

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 80–103
https://doi.org/10.1214/21-AOAS1503

SUBGROUP-EFFECTS MODELS FOR THE ANALYSIS OF PERSONAL TREATMENT EFFECTS

BY LING ZHOU[1,a], SHIQUAN SUN[2,b], HAODA FU[3,c] AND PETER X.-K. SONG[4,d]

[1]*Center for Statistical Research, Southwestern University of Finance and Economics,* [a]*zhouling@swufe.edu.cn*
[2]*School of Public Health, Xi'an Jiaotong University,* [b]*sqsunsph@xjtu.edu.cn*
[3]*Eli Lilly and Company,* [c]*fu_haoda@lilly.com*
[4]*Department of Biostatistics, University of Michigan,* [d]*pxsong@umich.edu*

The emerging field of precision medicine is transforming statistical analysis from the classical paradigm of population-average treatment effects into that of personal treatment effects. This new scientific mission has called for adequate statistical methods to assess heterogeneous covariate effects in regression analysis. This paper focuses on a subgroup analysis that consists of two primary analytic tasks: identification of treatment effect subgroups and individual group memberships, and statistical inference on treatment effects by subgroup. We propose an approach to synergizing supervised clustering analysis via alternating direction method of multipliers (ADMM) algorithm and statistical inference on subgroup effects via expectation-maximization (EM) algorithm. Our proposed procedure, termed as hybrid operation for subgroup analysis (HOSA), enjoys computational speed and numerical stability with interpretability and reproducibility. We establish key theoretical properties for both proposed clustering and inference procedures. Numerical illustration includes extensive simulation studies and analyses of motivating data from two randomized clinical trials to learn subgroup treatment effects.

REFERENCES

ACKERMAN, S. (1992). *Discovering the Brain*. National Academies Press, Washington.

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

BELLINGER, D., STILES, K. and NEEDLEMAN, H. (1992). Low-level lead exposure, intelligence and academic achievement: A long-term follow-up study. *Pediatrics* **90** 855–861.

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.

BOIVIN, M. and GIORDANI, B. (1995). A risk evaluation of the neuropsychological effects of childhood lead toxicity. *Dev. Neuropsychol.* **11** 157–180.

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.

CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning* 1040–1048.

CHEN, J., LI, P. and FU, Y. (2012). Inference on the order of a normal mixture. *J. Amer. Statist. Assoc.* **107** 1096–1105.

CHEN, Y., YI, X. and CARAMANIS, C. (2018). Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. *IEEE Trans. Inf. Theory* **64** 1738–1766. MR3766312 https://doi.org/10.1109/TIT.2017.2773474

CHI, E. C. and LANGE, K. (2015). Splitting methods for convex clustering. *J. Comput. Graph. Statist.* **24** 994–1013. MR3432926 https://doi.org/10.1080/10618600.2014.948181

Key words and phrases. ADMM algorithm, EM algorithm, maximum likelihood, precision medicine, supervised clustering.

CRAWFORD, S. L. (1994). An application of the Laplace method to finite mixture distributions. *J. Amer. Statist. Assoc.* **89** 259–267. MR1266298

DASGUPTA, A. and RAFTERY, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J. Amer. Statist. Assoc.* **93** 294–302.

DEB, P. and HOLMES, A. M. (2000). Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models. *Health Econ.* **9** 475–489.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DOBBIN, K. and SIMON, R. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8** 101–117.

ELHAMIFAR, E. and VIDAL, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 2765–2781.

ETTINGER, A., HU, H. and HERNANDEZ-AVILA, M. (2007). Dietary calcium supplementation to lower blood lead levels in pregnancy and lactation. *J. Nutr. Biochem* **18** 172–178.

ETTINGER, A., LAMADRID-FIGUEROA, H., TELLEZ-ROJO, M., MERCADO-GARCIA, A., PETERSON, K., SCHWARTZ, J., HU, H. and HERNANDEZ-AVILA, M. (2009). Effect of calcium supplementation on blood lead levels in pregnancy: A randomized placebo-controlled trial. *Environ. Health Perspect.* **117** 26–31.

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

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer, New York. MR2265601

GABAY, D. and MERCIER, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* **2** 17–40.

GLOWINSKI, R. (2014). On alternating direction methods of multipliers: A historical perspective. In *Modeling, Simulation and Optimization for Science and Technology. Comput. Methods Appl. Sci.* **34** 59–82. Springer, Dordrecht. MR3330832 https://doi.org/10.1007/978-94-017-9054-3_4

GRÜN, B. and LEISCH, F. (2007). Applications of finite mixtures of regression models. Available at http://cran.r-project.org/web/packages/flexmix/vignettes/regression-examples.pdf.

GUNTER, L., ZHU, J. and MURPHY, S. A. (2011). Variable selection for qualitative interactions. *Stat. Methodol.* **8** 42–55. MR2741508 https://doi.org/10.1016/j.stamet.2009.05.003

HU, H., TÉLLEZ-ROJO, M., BELLINGER, D., SMITH, D., ETTINGER, A., LAMADRID-FIGUEROA, H., SCHWARTZ, J., SCHNAAS, L., MERCADO-GARCIA, A. et al. (2006). Fetal lead exposure at each stage of pregnancy as a predictor of infant mental development. *Environ. Health Perspect.* **114** 1730–1735.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87. https://doi.org/10.1162/neco.1991.3.1.79

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. MR1769735

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

MCLACHLAN, G. J., LEE, S. X. and RATHNAYAKE, S. I. (2019). Finite mixture models. *Annu. Rev. Stat. Appl.* **6** 355–378. MR3939525 https://doi.org/10.1146/annurev-statistics-031017-100325

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

MCLACHLAN, G. J. and RATHNAYAKE, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **4** 341–355.

MIHIĆ, K., ZHU, M. and YE, Y. (2021). Managing randomization in the multi-block alternating direction method of multipliers for quadratic optimization. *Math. Program. Comput.* **13** 339–413. MR4266928 https://doi.org/10.1007/s12532-020-00192-5

MOTA, J. F. C., XAVIER, J. M. F., AGUIAR, P. M. Q. and PÜSCHEL, M. (2013). D-ADMM: A communication-efficient distributed algorithm for separable optimization. *IEEE Trans. Signal Process.* **61** 2718–2723. MR3053838 https://doi.org/10.1109/TSP.2013.2254478

MUTHÉN, B. and SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55** 463–469.

NORMAN, J., POLITZ, D. and POLITZ, L. (2009). Hyperparathyroidism during pregnancy and the effect of rising calcium on pregnancy loss: A call for earlier intervention. *Clin. Endocrinol.* **71** 104–109.

PERNG, W., TAMAYO-ORTIZ, M., TANG, L., SANCHEZ, B., CANTORAL, A., MEEKER, J., DOLINOY, D., ROBERTS, E., MIER, A. et al. (2019). The Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) Project. *British Med. J. Open* **9** e030427.

PIMENTEL-ALARCÓN, D., BALZANO, L., MARCIA, R., NOWAK, R. and WILLETT, R. (2017). Mixture regression as subspace clustering. In *Sampling Theory and Applications* (*SampTA*), 2017 *International Conference on* 456–459. IEEE, New York.

PROUST, C. and JACQMIN-GADDA, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput. Methods Programs Biomed.* **78** 165–173.
PROUST-LIMA, C., PHILIPPS, V. and LIQUET, B. (2017). Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *J. Stat. Softw.* **78** 1–56.
REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930 https://doi.org/10.1137/1026034
ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121 https://doi.org/10.2307/2965553
SANCHEZ, B., WU, M., SONG, P. and WANG, W. (2016). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **17** 722–736.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics* 1223–1231.
SOLTANOLKOTABI, M., ELHAMIFAR, E. and CANDÈS, E. J. (2014). Robust subspace clustering. *Ann. Statist.* **42** 669–699. MR3210983 https://doi.org/10.1214/13-AOS1199
SUN, R., LUO, Z.-Q. and YE, Y. (2015). On the expected convergence of randomly permuted ADMM. Preprint. Available at arXiv:1503.06387.
TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Stat.* **32** 244–248. MR0120677 https://doi.org/10.1214/aoms/1177705155
TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x
VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91** 217–221.
VIELE, K. and TONG, B. (2002). Modeling with mixtures of linear regressions. *Stat. Comput.* **12** 315–330. MR1951705 https://doi.org/10.1023/A:1020779827503
WEI, S. and KOSOROK, M. R. (2013). Latent supervised learning. *J. Amer. Statist. Assoc.* **108** 957–970. MR3174676 https://doi.org/10.1080/01621459.2013.789695
WONG, C. H., SIAH, K. W. and LO, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics* **20** 273–286. MR3922133 https://doi.org/10.1093/biostatistics/kxx069
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
XU, W. and HEDEKER, D. (2001). A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *J. Biopharm. Statist.* **11** 253–273.
YI, X., CARAMANIS, C. and SANGHAVI, S. (2014). Alternating minimization for mixed linear regression. In *International Conference on Machine Learning* 613–621.
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, Y., CHEN, X., ZHOU, D. and JORDAN, M. I. (2016). Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *J. Mach. Learn. Res.* **17** Paper No. 102, 44. MR3543508
ZHONG, K., JAIN, P. and DHILLON, I. S. (2016). Mixed linear regression with multiple components. In *Advances in Neural Information Processing Systems* 2190–2198.
ZHOU, L., SUN, S., FU, H. and SONG, P. X. (2022). Supplement to "Subgroup-Effects Models for the Analysis of Personal Treatment Effects." https://doi.org/10.1214/21-AOAS1503SUPPA, https://doi.org/10.1214/21-AOAS1503SUPPB

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 104–123
https://doi.org/10.1214/21-AOAS1488

INFERENCE IN BAYESIAN ADDITIVE VECTOR AUTOREGRESSIVE TREE MODELS

BY FLORIAN HUBER[1,a] AND LUCA ROSSINI[2,b]

[1]*Department of Economics, University of Salzburg,* [a]*florian.huber@sbg.ac.at*
[2]*Department of Economics, Management and Quantitative Methods, University of Milan,* [b]*luca.rossini@unimi.it*

Vector autoregressive (VAR) models assume linearity between the endogenous variables and their lags. This assumption might be overly restrictive and could have a deleterious impact on forecasting accuracy. As a solution we propose combining VAR with Bayesian additive regression tree (BART) models. The resulting Bayesian additive vector autoregressive tree (BAVART) model is capable of capturing arbitrary nonlinear relations between the endogenous variables and the covariates without much input from the researcher. Since controlling for heteroscedasticity is key for producing precise density forecasts, our model allows for stochastic volatility in the errors. We apply our model to two datasets. The first application shows that the BAVART model yields highly competitive forecasts of the U.S. term structure of interest rates. In a second application we estimate our model using a moderately sized Eurozone dataset to investigate the dynamic effects of uncertainty on the economy.

REFERENCES

AASTVEIT, K. A., NATVIK, G. J. and SOLA, S. (2017). Economic uncertainty and the influence of monetary policy. *J. Int. Money Financ.* **76** 50–67.

ALESSANDRI, P. and MUMTAZ, H. (2017). Financial conditions and density forecasts for US output and inflation. *Rev. Econ. Dyn.* **24** 66–78.

ALESSANDRI, P. and MUMTAZ, H. (2019). Financial regimes and uncertainty shocks. *J. Monet. Econ.* **101** 31–46.

AUERBACH, A. J. and GORODNICHENKO, Y. (2012). Measuring the output responses to fiscal policy. *Am. Econ. J. Econ. Policy* **4** 1–27.

BARNICHON, R. and MATTHES, C. (2018). Functional approximation of impulse responses. *J. Monet. Econ.* **99** 41–55.

BASSETTI, F., CASARIN, R. and LEISEN, F. (2014). Beta-product dependent Pitman–Yor processes for Bayesian inference. *J. Econometrics* **180** 49–72. MR3188911 https://doi.org/10.1016/j.jeconom.2014.01.007

BILLIO, M., CASARIN, R. and ROSSINI, L. (2019). Bayesian nonparametric sparse VAR models. *J. Econometrics* **212** 97–115. MR3994009 https://doi.org/10.1016/j.jeconom.2019.04.022

BLOOM, N. (2009). The impact of uncertainty shocks. *Econometrica* **77** 623–685. MR2531358 https://doi.org/10.3982/ECTA6248

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

CAGGIANO, G., CASTELNUOVO, E. and GROSHENNY, N. (2014). Uncertainty shocks and unemployment dynamics in US recessions. *J. Monet. Econ.* **67** 78–92.

CAGGIANO, G., CASTELNUOVO, E. and NODARI, G. (2021). Uncertainty and monetary policy in good and bad times. *J. Appl. Econometrics*. To appear.

CAGGIANO, G., CASTELNUOVO, E. and PELLEGRINO, G. (2017). Estimating the real effects of uncertainty shocks at the zero lower bound. *Eur. Econ. Rev.* **100** 257–272.

CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2018). Measuring uncertainty and its impact on the economy. *Rev. Econ. Stat.* **100** 799–815.

CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *J. Econometrics* **212** 137–154. MR3994011 https://doi.org/10.1016/j.jeconom.2019.04.024

CARRIERO, A., KAPETANIOS, G. and MARCELLINO, M. (2012). Forecasting government bond yields with large Bayesian vector autoregressions. *J. Bank. Financ.* **36** 2026–2047.

Key words and phrases. Bayesian additive regression trees, BAVART, decision trees, nonparametric regression, vector autoregressions.

CARRIERO, A., CHAN, J. C. C., CLARK, T. E. and MARCELLINO, M. (2021). Corrigendum to: Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. Working paper.
CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80. PMLR.
CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.
CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2002). Bayesian treed models. *Mach. Learn.* **48** 299–320.
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. and RAVAZZOLO, F. (2015). Macroeconomic forecasting performance under alternative specifications of time-varying volatility. *J. Appl. Econometrics* **30** 551–575. MR3358636 https://doi.org/10.1002/jae.2379
CRESPO CUARESMA, J. C., HUBER, F. and ONORANTE, L. (2020). Fragility and the effect of international uncertainty shocks. *J. Int. Money Financ.* **108** 102151.
DIEBOLD, F. X. and LI, C. (2006). Forecasting the term structure of government bond yields. *J. Econometrics* **130** 337–364. MR2211798 https://doi.org/10.1016/j.jeconom.2005.03.005
DOAN, T., LITTERMAN, R. and SIMS, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Rev.* **3** 1–100.
DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* **34** 43–68. MR3938963 https://doi.org/10.1214/18-STS667
FERRARA, L. and GUÉRIN, P. (2018). What are the macroeconomic effects of high-frequency uncertainty shocks? *J. Appl. Econometrics* **33** 662–679. MR3850753 https://doi.org/10.1002/jae.2624
FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055 https://doi.org/10.1006/jcss.1997.1504
FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451
GEFANG, D. and STRACHAN, R. (2010). Nonlinear impacts of international business cycles on the U.K.—a Bayesian smooth transition VAR approach. *Stud. Nonlinear Dyn. Econom.* **14** Art. 2, 33. MR2585173 https://doi.org/10.2202/1558-3708.1677
GIESECK, A. and LARGENT, Y. (2016). The impact of macroeconomic uncertainty on activity in the euro area. *Rev. Econ.* **67** 25–52.
GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statist. Assoc.* **103** 1119–1130. MR2528830 https://doi.org/10.1198/016214508000000689
GRANGER, C. W. and TERASVIRTA, T. (1993). Modelling non-linear economic relationships. OUP Catalogue.
GREEN, D. P. and KERN, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* **76** 491–511.
GÜRKAYNAK, R. S., SACK, B. and WRIGHT, J. H. (2007). The US treasury yield curve: 1961 to the present. *J. Monet. Econ.* **54** 2291–2304.
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
HE, J., YALOV, S. and HAHN, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *The* 22*nd International Conference on Artificial Intelligence and Statistics* 1130–1138. PMLR.
HERNÁNDEZ, B., RAFTERY, A. E., PENNINGTON, S. R. and PARNELL, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Stat. Comput.* **28** 869–890. MR3766048 https://doi.org/10.1007/s11222-017-9767-1
HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. Supplementary material available online. MR2816546 https://doi.org/10.1198/jcgs.2010.08162
HILL, J., LINERO, A. and MURRAY, J. (2020). Bayesian additive regression trees: A review and look forward. *Annu. Rev. Stat. Appl.* **7** 251–278. MR4104193 https://doi.org/10.1146/annurev-statistics-031219-041110
HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G. (2010). *Bayesian Nonparametrics*. Cambridge Univ. Press, Cambridge.
HUBER, F., KOOP, G. and ONORANTE, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *J. Bus. Econom. Statist.* **39** 669–683. MR4272927 https://doi.org/10.1080/07350015.2020.1713796

HUBER, F. and ZÖRNER, T. O. (2019). Threshold cointegration in international exchange rates: A Bayesian approach. *Int. J. Forecast.* **35** 458–473.

HUBER, F., KOOP, G., ONORANTE, L., PFARRHOFER, M. and SCHREINER, J. (2021). Nowcasting in a pandemic using non-parametric mixed frequency VARs. *J. Econometrics*. To appear.

JACKSON, L. E., KLIESEN, K. L. and OWYANG, M. T. (2020). The nonlinear effects of uncertainty shocks. *Stud. Nonlinear Dyn. Econom.* **24** 20190024, 19. MR4145962 https://doi.org/10.1515/snde-2019-0024

JURADO, K., LUDVIGSON, S. C. and NG, S. (2015). Measuring uncertainty. *Am. Econ. Rev.* **105** 1177–1216.

KALLI, M. and GRIFFIN, J. E. (2018). Bayesian nonparametric vector autoregressive models. *J. Econometrics* **203** 267–282. MR3770826 https://doi.org/10.1016/j.jeconom.2017.11.009

KAPELNER, A. and BLEICH, J. (2015). Prediction with missing data via Bayesian additive regression trees. *Canad. J. Statist.* **43** 224–239. MR3353381 https://doi.org/10.1002/cjs.11248

KASTNER, G. and FRÜHWIRTH-SCHNATTER, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Comput. Statist. Data Anal.* **76** 408–423. MR3209449 https://doi.org/10.1016/j.csda.2013.01.002

KASTNER, G. and HUBER, F. (2020). Sparse Bayesian vector autoregressions in huge dimensions. *J. Forecast.* **39** 1142–1165. MR4161021 https://doi.org/10.1002/for.2680

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

KOOP, G., KOROBILIS, D. and PETTENUZZO, D. (2019). Bayesian compressed vector autoregressions. *J. Econometrics* **210** 135–154. MR3944767 https://doi.org/10.1016/j.jeconom.2018.11.009

KOOP, G., PESARAN, M. H. and POTTER, S. M. (1996). Impulse response analysis in nonlinear multivariate models. *J. Econometrics* **74** 119–147. MR1409037 https://doi.org/10.1016/0304-4076(95)01753-4

KRUEGER, R., BANSAL, P. and BUDDHAVARAPU, P. (2020). A new spatial count data model with Bayesian additive regression trees for accident hot spot identification. *Accident Anal. Prev.* **144** 105623. https://doi.org/10.1016/j.aap.2020.105623

LAKSHMINARAYANAN, B., ROY, D. M. and TEH, Y. W. (2014). Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) **27** 3140–3148. Curran Associates.

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

LITTERMAN, R. B. (1986). Forecasting with Bayesian vector autoregressions—five years of experience. *J. Bus. Econom. Statist.* **4** 25–38.

MAKALIC, E. and SCHMIDT, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Process. Lett.* **23** 179–182.

MUMTAZ, H. and THEODORIDIS, K. (2018). The changing transmission of uncertainty shocks in the U.S. *J. Bus. Econom. Statist.* **36** 239–252. MR3790211 https://doi.org/10.1080/07350015.2016.1147357

NELSON, C. R. and SIEGEL, A. F. (1987). Parsimonious modeling of yield curves. *J. Bus.* **60** 473–489.

PACCAGINI, A. and COLOMBO, V. (2020). The asymmetric effect of uncertainty shocks. CAMA Working Paper, 72/2020.

PLAGBORG-MØLLER, M. (2019). Bayesian inference on structural impulse response functions. *Quant. Econ.* **10** 145–184. MR3915253 https://doi.org/10.3982/QE926

PRATOLA, M. T., CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *J. Comput. Graph. Statist.* **29** 405–417. MR4116052 https://doi.org/10.1080/10618600.2019.1677243

PRÜSER, J. (2019). Forecasting with many predictors using Bayesian additive regression trees. *J. Forecast.* **38** 621–631. MR4021441 https://doi.org/10.1002/for.2587

RAMEY, V. A. and ZUBAIRY, S. (2017). Government spending multipliers in good times and in bad: Evidence from us historical data. *J. Polit. Econ.* **126** 850–901.

ROY, D. M. and TEH, Y. W. (2009). The mondrian process. In *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) **21** 1377–1384. Curran Associates.

SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.

SIMS, C. A. and ZHA, T. (1998). Bayesian methods for dynamic multivariate models. *Internat. Econom. Rev.* **39** 949–968.

WALDMANN, P. (2016). Genome-wide prediction using Bayesian additive regression trees. *Genet. Sel. Evol.* **48** 1–12.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 124–143
https://doi.org/10.1214/21-AOAS1489

A FLEXIBLE BAYESIAN FRAMEWORK TO ESTIMATE AGE- AND CAUSE-SPECIFIC CHILD MORTALITY OVER TIME FROM SAMPLE REGISTRATION DATA

BY AUSTIN E. SCHUMACHER[1,a], TYLER H. MCCORMICK[2,c], JON WAKEFIELD[3,d], YUE CHU[4,e], JAMIE PERIN[5,f], FRANCISCO VILLAVICENCIO[5,g], NOAH SIMON[1,b] AND LI LIU[6,h]

[1]*Department of Biostatistics, University of Washington,* [a]*aeschuma@uw.edu,* [b]*nrsimon@uw.edu*
[2]*Departments of Statistics and Sociology, University of Washington,* [c]*tylermc@uw.edu*
[3]*Departments of Biostatistics and Statistics, University of Washington,* [d]*jonno@uw.edu*
[4]*Department of Sociology, The Ohio State University,* [e]*chu.282@osu.edu*
[5]*Department of International Health, Johns Hopkins Bloomberg School of Public Health,* [f]*jperin@jhu.edu,* [g]*fvillav1@jhu.edu*
[6]*Departments of Population, Family and Reproductive Health and International Health, Johns Hopkins Bloomberg School of Public Health,* [h]*lliu26@jhu.edu*

In order to implement disease-specific interventions in young age groups, policy makers in low- and middle-income countries require timely and accurate estimates of age- and cause-specific child mortality. High-quality data is not available in settings where these interventions are most needed, but there is a push to create sample registration systems that collect detailed mortality information. Current methods that estimate mortality from this data employ multistage frameworks without rigorous statistical justification that separately estimate all-cause and cause-specific mortality and are not sufficiently adaptable to capture important features of the data. We propose a flexible Bayesian modeling framework to estimate age- and cause-specific child mortality from sample registration data. We provide a theoretical justification for the framework, explore its properties via simulation, and use it to estimate mortality trends using data from the Maternal and Child Health Surveillance System in China.

REFERENCES

ABDULLAH, S., ADAZU, K., MASANJA, H., DIALLO, D., HODGSON, A., ILBOUDO-SANOGO, E., NHACOLO, A., OWUSU-AGYEI, S., THOMPSON, R. et al. (2007). Patterns of age-specific mortality in children in endemic areas of sub-Saharan Africa. *Am. J. Trop. Med. Hyg.* **77** 99–105.

ABOUZAHR, C., DE SAVIGNY, D., MIKKELSEN, L., SETEL, P. W., LOZANO, R. and LOPEZ, A. D. (2015a). Towards universal civil registration and vital statistics systems: The time is now. *Lancet* **386** 1407–1418.

ABOUZAHR, C., DE SAVIGNY, D., MIKKELSEN, L., SETEL, P. W., LOZANO, R., NICHOLS, E., NOTZON, F. and LOPEZ, A. D. (2015b). Civil registration and vital statistics: Progress in the data revolution for counting and accountability. *Lancet* **386** 1373–1385.

ALKEMA, L. and NEW, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline Bias-reduction model. *Ann. Appl. Stat.* **8** 2122–2149. MR3292491 https://doi.org/10.1214/14-AOAS768

APONTE, J. J., SCHELLENBERG, D., EGAN, A., BRECKENRIDGE, A., CARNEIRO, I., CRITCHLEY, J., DANQUAH, I., DODOO, A., KOBBE, R. et al. (2009). Efficacy and safety of intermittent preventive treatment with sulfadoxine-pyrimethamine for malaria in African infants: A pooled analysis of six randomised, placebo-controlled trials. *Lancet* **374** 1533–1542.

BCHIR, A., BHUTTA, Z., BINKA, F., BLACK, R., BRADSHAW, D., GARNETT, G., HAYASHI, K., JHA, P., PETO, R. et al. (2006). Better health statistics are possible. *Lancet* **367** 190–193.

BENNETT, J. and WAKEFIELD, J. (2001). Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics* **57** 803–812. MR1863449 https://doi.org/10.1111/j.0006-341X.2001.00803.x

Key words and phrases. Bayesian inference, sample registration system, child mortality, cause-specific mortality.

BOERMA, J. T. (2013). Public health information needs in districts. *BMC Health Serv. Res.* **13** S12.

BOERMA, J. T. and STANSFIELD, S. K. (2007). Health statistics now: Are we making the right investments? *Lancet* **369** 779–786.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CHI, E. M. and REINSEL, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *J. Amer. Statist. Assoc.* **84** 452–459. MR1010333

CLARK, S. J., SETEL, P. and LI, Z. (2019). Verbal autopsy in civil registration and vital statistics: The symptom-cause information archive. Preprint. Available at arXiv:1910.00405.

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016). 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

DATTA, A., BANERJEE, S., HODGES, J. S. and GAO, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. *Bayesian Anal.* **14** 1221–1244. MR4044851 https://doi.org/10.1214/19-BA1177

DESAI, N., ALEKSANDROWICZ, L., MIASNIKOF, P., LU, Y., LEITAO, J., BYASS, P., TOLLMAN, S., MEE, P., ALAM, D. et al. (2014). Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC Med.* **12** 20.

FRIBERG, I. K., KINNEY, M. V., LAWN, J. E., KERBER, K. J., ODUBANJO, M. O., BERGH, A.-M., WALKER, N., WEISSMAN, E., CHOPRA, M. et al. (2010). Sub-Saharan Africa's mothers, newborns, and children: How many lives could be saved with targeted health interventions? *PLoS Med.* **7** e1000295.

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

GLASS, R. I., GUTTMACHER, A. E. and BLACK, R. E. (2012). Ending preventable child death in a generation. *J. Am. Med. Assoc.* **308** 141–142.

HE, C., LIU, L., CHU, Y., PERIN, J., DAI, L., LI, X., MIAO, L., KANG, L., LI, Q. et al. (2017). National and subnational all-cause and cause-specific child mortality in China, 1996–2015: A systematic analysis with implications for the sustainable development goals. *The Lancet Global Health* **5** e186–e197.

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

HELD, L., SCHRÖDLE, B. and RUE, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures* 91–110. Physica-Verlag/Springer, Heidelberg. MR2664630 https://doi.org/10.1007/978-3-7908-2413-1_6

HOLFORD, T. R. (1976). Life tables with concomitant information. *Biometrics* 587–597.

HOLFORD, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* **36** 299–305.

JHA, P. (2012). Counting the dead is one of the world's best investments to reduce premature mortality. *Hypothesis* **10** e3.

KEENAN, J. D., BAILEY, R. L., WEST, S. K., ARZIKA, A. M., HART, J., WEAVER, J., KALUA, K., MRANGO, Z., RAY, K. J. et al. (2018). Azithromycin to reduce childhood mortality in sub-Saharan Africa. *N. Engl. J. Med.* **378** 1583–1592.

KELLER, J. P., OLIVES, C., KIM, S.-Y., SHEPPARD, L., SAMPSON, P. D., SZPIRO, A. A., ORON, A. P., LINDSTRÖM, J., VEDAL, S. and KAUFMAN, J. D. (2015). A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environ. Health Perspect.* **123** 301–309.

LAIRD, N. and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76** 231–240. MR0624329

LEE, J. Y., GREEN, P. J. and RYAN, L. M. (2017). On the "Poisson trick" and its extensions for fitting multinomial regression models. Preprint. Available at arXiv:1707.08538.

LI, Z., HSIAO, Y., GODWIN, J., MARTIN, B. D., WAKEFIELD, J., CLARK, S. J., WITH SUPPORT FROM THE UNITED NATIONS INTER-AGENCY GROUP FORCHILD MORTALITY ESTIMATION and ITS TECHNICAL ADVISORYGROUP (2019). Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PLoS ONE* **14** e0210645.

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

LIU, S., WU, X., LOPEZ, A. D., WANG, L., CAI, Y., PAGE, A., YIN, P., LIU, Y., LI, Y. et al. (2016a). An integrated national mortality surveillance system for death registration and mortality surveillance, China. *Bull. World Health Organ.* **94** 46–57.

LIU, L., OZA, S., HOGAN, D., CHU, Y., PERIN, J., ZHU, J., LAWN, J. E., COUSENS, S., MATHERS, C. et al. (2016b). Global, regional, and national causes of under-5 mortality in 2000–15: An updated systematic analysis with implications for the sustainable development goals. *Lancet* **388** 3027–3035.

MAHAPATRA, P. (2010). An overview of the sample registration system in India. In *Prince Mahidol Award Conference & Global Health Information Forum* 27–30.

MAHER, D., BIRARO, S., HOSEGOOD, V., ISINGO, R., LUTALO, T., MUSHATI, P., NGWIRA, B., NYIRENDA, M., TODD, J. et al. (2010). Translating global health research aims into action: The example of the ALPHA network. *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

MIKKELSEN, L., PHILLIPS, D. E., ABOUZAHR, C., SETEL, P. W., DE SAVIGNY, D., LOZANO, R. and LOPEZ, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *Lancet* **386** 1395–1406.

MURRAY, C. J., LOZANO, R., FLAXMAN, A. D., SERINA, P., PHILLIPS, D., STEWART, A., JAMES, S. L., VAHDATPOUR, A., ATKINSON, C. et al. (2014). Using verbal autopsy to measure causes of death: The comparative performance of existing methods. *BMC Med.* **12** 5.

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

O'BRIEN, K. L., WOLFSON, L. J., WATT, J. P., HENKLE, E., DELORIA-KNOLL, M., MCCALL, N., LEE, E., MULHOLLAND, K., LEVINE, O. S. et al. (2009). Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: Global estimates. *Lancet* **374** 893–902.

PENNY, M. A., VERITY, R., BEVER, C. A., SAUBOIN, C., GALACTIONOVA, K., FLASCHE, S., WHITE, M. T., WENGER, E. A., VAN DE VELDE, N. et al. (2016). Public health impact and cost-effectiveness of the RTS,S/AS01 malaria vaccine: A systematic comparison of predictions from four mathematical models. *Lancet* **387** 367–375.

PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. MR3075338 https://doi.org/10.1214/12-STS395

PHILLIPS, D. E., ABOUZAHR, C., LOPEZ, A. D., MIKKELSEN, L., DE SAVIGNY, D., LOZANO, R., WILMOTH, J. and SETEL, P. W. (2015). Are well functioning civil registration and vital statistics systems associated with better health outcomes? *Lancet* **386** 1386–1394.

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

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON JR, A. V., FLOURNOY, N., FAREWELL, V. T. and BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* 541–554.

R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

RAO, C., SOEMANTRI, S., DJAJA, S., ADAIR, T., WIRYAWAN, Y., PANGARIBUAN, L., IRIANTO, J., KOSEN, S., LOPEZ, A. D. et al. (2010). Mortality in central Java: Results from the Indonesian mortality registration system strengthening project. *BMC Res. Notes* **3** 325.

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

ROBERTS, D. R., BAHN, V., CIUTI, S., BOYCE, M. S., ELITH, J., GUILLERA-ARROITA, G., HAUENSTEIN, S., LAHOZ-MONFORT, J. J., SCHRÖDER, B. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40** 913–929.

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

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

SCHUMACHER, A. E., MCCORMICK, T. H., WAKEFIELD, J., CHU, Y., PERIN, J., VILLAVICENCIO, F., SIMON, N. and LIU, L. (2022). Supplement to "A flexible Bayesian framework to estimate age- and cause-specific child mortality over time from sample registration data." https://doi.org/10.1214/21-AOAS1489SUPPA, https://doi.org/10.1214/21-AOAS1489SUPPB

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

SMITH, M. S. and KHALED, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *J. Amer. Statist. Assoc.* **107** 290–303. MR2949360 https://doi.org/10.1080/01621459.2011.644501

SNOW, R. W., OMUMBO, J. A., LOWE, B., MOLYNEUX, C. S., OBIERO, J.-O., PALMER, A., WEBER, M. W., PINDER, M., NAHLEN, B. et al. (1997). Relation between severe malaria morbidity in children and level of Plasmodium falciparum transmission in Africa. *Lancet* **349** 1650–1654.

SOLEMAN, N., CHANDRAMOHAN, D. and SHIBUYA, K. (2006). Verbal autopsy: Current practices and challenges. *Bull. World Health Organ.* **84** 239–245.

SPECKMAN, P. L. and SUN, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90** 289–302. MR1986647 https://doi.org/10.1093/biomet/90.2.289

UNITED NATIONS (2015). Transforming our world: The 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September 2015.

UNITED NATIONS INTER-AGENCY GROUP FOR CHILD MORTALITY ESTIMATION (2020). Levels & Trends in Child Mortality 2020, Technical Report, United Nations Children's Fund.

UNITED NATIONS POPULATION DIVISION (2019). World Population Prospects, Technical Report, Dept. of International Economic and Social Affairs.

VOS, T., LIM, S. S., ABBAFATI, C., ABBAS, K. M., ABBASI, M., ABBASIFARD, M., ABBASI-KANGEVARI, M., ABBASTABAR, H., ABD-ALLAH, F. et al. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396** 1204–1222.

WAKEFIELD, J., FUGLSTAD, G.-A., RIEBLER, A., GODWIN, J., WILSON, K. and CLARK, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Stat. Methods Med. Res.* **28** 2614–2634. MR4000184 https://doi.org/10.1177/0962280218767988

WALKER, N., BRYCE, J. and BLACK, R. E. (2007). Interpreting health statistics for policymaking: The story behind the headlines. *Lancet* **369** 956–963.

WALKER, C. L. F., RUDAN, I., LIU, L., NAIR, H., THEODORATOU, E., BHUTTA, Z. A., O'BRIEN, K. L., CAMPBELL, H. and BLACK, R. E. (2013). Global burden of childhood pneumonia and diarrhoea. *Lancet* **381** 1405–1416.

WANG, H., ABBAS, K. M., ABBASIFARD, M., ABBASI-KANGEVARI, M., ABBASTABAR, H., ABD-ALLAH, F., ABDELALIM, A., ABOLHASSANI, H., ABREU, L. G. et al. (2020). Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: A comprehensive demographic analysis for the Global Burden of Disease Study 2019. *Lancet* **396** 1160–1203.

WHELDON, M. C., RAFTERY, A. E., CLARK, S. J. and GERLAND, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *J. Amer. Statist. Assoc.* **108** 96–110. MR3174605 https://doi.org/10.1080/01621459.2012.737729

WHO COLLABORATIVE STUDY TEAM ON THE ROLE OF BREASTFEEDING ON THE PREVENTION OF INFANT MORTALITY (2001). Effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: A pooled analysis. *Lancet* **355** 451–455.

YANG, G., HU, J., RAO, K. Q., MA, J., RAO, C. and LOPEZ, A. D. (2005). Mortality registration and surveillance in China: History, current situation and challenges. *Popul. Health Metr.* **3** 3.

YOU, D., HUG, L., EJDEMYR, S., IDELE, P., HOGAN, D., MATHERS, C., GERLAND, P., NEW, J. R., ALKEMA, L. et al. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *Lancet* **386** 2275–2286.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 144–168
https://doi.org/10.1214/21-AOAS1494

BAYESIAN NONPARAMETRIC MULTIVARIATE SPATIAL MIXTURE MIXED EFFECTS MODELS WITH APPLICATION TO AMERICAN COMMUNITY SURVEY SPECIAL TABULATIONS

BY RYAN JANICKI[1,a], ANDREW M. RAIM[1,b], SCOTT H. HOLAN[2,3,d] AND JERRY J. MAPLES[1,c]

[1]*Center for Statistical Research and Methodology, U.S. Census Bureau,* [a]*ryan.janicki@census.gov,* [b]*andrew.raim@census.gov,* [c]*jerry.j.maples@census.gov*

[2]*Department of Statistics, University of Missouri,* [d]*holans@missouri.edu*

[3]*Office of the Associate Director for Research and Methodology, U.S. Census Bureau*

Leveraging multivariate spatial dependence to improve the precision of estimates using American Community Survey data and other sample survey data has been a topic of recent interest among data users and federal statistical agencies. One strategy is to use a multivariate spatial mixed effects model with a Gaussian observation model and latent Gaussian process model. In practice, this works well for a wide range of tabulations. Nevertheless, in situations in which the data exhibit heterogeneity within or across geographies, and/or there is sparsity in the data, the Gaussian assumptions may be problematic and lead to underperformance. To remedy these situations, we propose a multivariate hierarchical Bayesian nonparametric mixed effects spatial mixture model to increase model flexibility. The number of clusters is chosen automatically in a data-driven manner. The effectiveness of our approach is demonstrated through a simulation study and motivating application of special tabulations for American Community Survey data.

REFERENCES

ABOWD, J. M. (2018). The US Census Bureau adopts differential privacy. In *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2867–2867.

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

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2016). A comparison of spatial predictors when datasets could be very large. *Stat. Surv.* **10** 100–131. MR3527662 https://doi.org/10.1214/16-SS115

BRADLEY, J. R., HOLAN, S. H. and WIKLE, C. K. (2015). Multivariate spatio-temporal 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. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Anal.* **13** 253–310. MR3773410 https://doi.org/10.1214/17-BA1069

BRADLEY, J. R., WIKLE, C. K. and HOLAN, S. H. (2017). Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 815–832. MR3641409 https://doi.org/10.1111/rssb.12179

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

DIGGLE, P. J., TAWN, J. A. and MOYEED, R. A. (1998). Model-based geostatistics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 299–350. MR1626544 https://doi.org/10.1111/1467-9876.00113

DUAN, J. A., GUINDANI, M. and GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94** 809–825. MR2416794 https://doi.org/10.1093/biomet/asm071

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510 https://doi.org/10.1080/01621459.1995.10476550

Key words and phrases. American Community Survey, Dirichlet process, mixture models, nonparametric Bayes, small area estimation.

FAY, R. E. and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019 https://doi.org/10.1080/01621459.1979.10482505

FAY, R. and TRAIN, G. (1995). Aspects of survey and model based postcensal estimation of income and poverty characteristics for states and counties. In *Joint Statistical Meetings*: *Proceedings of the Section of Government Statistics* 154–159.

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

FERNÁNDEZ, C. and GREEN, P. J. (2002). Modelling spatially correlated data via mixtures: A Bayesian approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 805–826. MR1979388 https://doi.org/10.1111/1467-9868.00362

GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100** 1021–1035. MR2201028 https://doi.org/10.1198/016214504000002078

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

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford Univ. Press, New York. MR1380276

HIGHAM, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR0943997 https://doi.org/10.1016/0024-3795(88)90223-6

HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G. (2010). *Bayesian Nonparametrics* **28**. Cambridge Univ. Press, Cambridge.

HOSSAIN, M. M., LAWSON, A. B., CAI, B., CHOI, J., LIU, J. and KIRBY, R. S. (2013). Space-time stick-breaking processes for small area disease cluster estimation. *Environ. Ecol. Stat.* **20** 91–107. MR3039588 https://doi.org/10.1007/s10651-012-0209-0

HOSSEINPOURI, M. and KHALEDI, M. J. (2019). An area-specific stick breaking process for spatial data. *Statist. Papers* **60** 199–221. MR3905446 https://doi.org/10.1007/s00362-016-0833-0

HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275 https://doi.org/10.1111/j.1467-9868.2012.01041.x

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729 https://doi.org/10.1198/016214501750332758

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, Calif. MR0133191

JANICKI, R., RAIM, A. M., HOLAN, S. H. and MAPLES, J. (2022). Supplement to "Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American Community Survey special tabulations." https://doi.org/10.1214/21-AOAS1494SUPPA, https://doi.org/10.1214/21-AOAS1494SUPPB

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478 https://doi.org/10.1198/016214506000000492

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

KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711. MR0350950 https://doi.org/10.1214/aop/1176996898

KOTTAS, A. (2016). Bayesian nonparametric modeling for disease incidence data. In *Handbook of Spatial Epidemiology* 363–374. CRC Press/CRC, London.

MOLINA, I., NANDRAM, B. and RAO, J. N. K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *Ann. Appl. Stat.* **8** 852–885. MR3262537 https://doi.org/10.1214/13-AOAS702

MORAN, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37** 17–23. MR0035933 https://doi.org/10.1093/biomet/37.1-2.17

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

NEELON, B., GELFAND, A. E. and MIRANDA, M. L. (2014). A multivariate spatial mixture model for areal data: Examining regional differences in standardized test scores. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 737–761. MR3269410 https://doi.org/10.1111/rssc.12061

PORTER, A. T., HOLAN, S. H. and WIKLE, C. K. (2015). Bayesian semiparametric hierarchical empirical likelihood spatial models. *J. Statist. Plann. Inference* **165** 78–90. MR3350260 https://doi.org/10.1016/j.jspi.2015.04.002

QIU, Y. and MEI, J. (2019). RSpectra: Solvers for large-scale eigenvalue and SVD problems. R package version 0.16-0.

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

REICH, B. J. and FUENTES, M. (2015). Spatial Bayesian nonparametric methods. In *Nonparametric Bayesian Inference in Biostatistics*. *Front. Probab. Stat. Sci*. 347–357. Springer, Cham. MR3411028

R CORE TEAM (2019). R: A language and environment for statistical computing R Foundation for Statistical Computing.

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

STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan.

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

TORRIERI, N. (2014). American Community Survey design and methodology. Technical Report, United States Census Bureau.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 169–192
https://doi.org/10.1214/21-AOAS1444

SPARSE MATRIX LINEAR MODELS FOR STRUCTURED HIGH-THROUGHPUT DATA

BY JANE W. LIANG[1,a] AND ŚAUNAK SEN[2,b]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [a]*jwliang@harvard.edu*
[2]*Department of Preventive Medicine, University of Tennessee Health Science Center,* [b]*sen@uthsc.edu*

Recent technological advancements have led to the rapid generation of high-throughput biological data which can be used to address novel scientific questions in broad areas of research. These data can be thought of as a large matrix with covariates annotating both its rows and columns. Matrix linear models provide a convenient way for modeling such data. In many situations, sparse estimation of these models is desired. We present fast, general methods for fitting sparse matrix linear models to structured high-throughput data. We induce model sparsity using an L_1 penalty and consider the case when the response matrix and the covariate matrices are large. Due to data size, standard methods for estimation of these penalized regression models fail if the problem is converted to the corresponding univariate regression scenario. By leveraging matrix properties in the structure of our model, we develop several fast estimation algorithms (coordinate descent, FISTA and ADMM) and discuss their trade-offs. We evaluate our method's performance on simulated data, *E. coli* chemical genetic screening data and two *Arabidopsis* genetic datasets with multivariate responses. Our algorithms have been implemented in the Julia programming language and are available at https://github.com/senresearch/MatrixLMnet.jl.

REFERENCES

ÅGREN, J., OAKLEY, C. G., MCKAY, J. K., LOVELL, J. T. and SCHEMSKE, D. W. (2013). Genetic mapping of adaptation reveals fitness tradeoffs in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **110** 21077–21082.

ÅGREN, J., OAKLEY, C. G., LUNDEMO, S. and SCHEMSKE, D. W. (2016). Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping. Data from: Dryad Digital Repository. Available at https://doi.org/10.5061/dryad.77971.

ÅGREN, J., OAKLEY, C. G., LUNDEMO, S. and SCHEMSKE, D. W. (2017). Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping. *Evolution* **71** 550–564.

BABA, T., ARA, T., HASEGAWA, M., TAKAI, Y., OKUMURA, Y., BABA, M., DATSENKO, K. A., TOMITA, M., WANNER, B. L. et al. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* **2**. https://doi.org/10.1038/msb4100050

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527 https://doi.org/10.1137/080716542

BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25** 60–83.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. MR3605826 https://doi.org/10.1137/141000671

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

BROMAN, K. W., WU, H., SEN, Ś. and CHURCHILL, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19** 889–890.

BUTLAND, G., BABU, M., DÍAZ-MEJÍA, J. J., BOHDANA, F., PHANSE, S., GOLD, B., YANG, W., LI, J., GAGARINOVA, A. G. et al. (2008). eSGA: E. coli synthetic genetic array analysis. *Nat. Methods* **5** 789–795.

DOWLE, M. and SRINIVASAN, A. (2018). data.table: Extension of 'data.frame'. R package version 1.11.8.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. and SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica* **12** 111–139. MR1894191

Key words and phrases. FISTA, ADMM, proximal gradient algorithms, gradient descent, Julia, LASSO.

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

EKSTRØM, C. T. (2018). MESS: Miscellaneous Esoteric Statistical Scripts. R package version 0.5.2.

FLOREA, M. I. and VOROBYOV, S. A. (2017). A robust FISTA-like algorithm. In 2017 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) 4521–4525. IEEE.

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

FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710 https://doi.org/10.2307/1390712

GHADIMI, E., TEIXEIRA, A., SHAMES, I. and JOHANSSON, M. (2012). On the optimal step-size selection for the alternating direction method of multipliers. *IFAC Proceedings Volumes* **45** 139–144.

HOBBS, E. C., ASTARITA, J. L. and STORZ, G. (2010). Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in Escherichia coli: Analysis of a bar-coded mutant collection. *J. Bacteriol.* **192** 59–67. https://doi.org/10.1128/JB.00873-09

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

KIM, D. and FESSLER, J. A. (2018). Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). *SIAM J. Optim.* **28** 223–250. MR3755677 https://doi.org/10.1137/16M108940X

LIANG, J. W., NICHOLS, R. J. and SEN, Ś. (2019). Matrix linear models for high-throughput chemical genetic screens. *Genetics* **212** 1063–1073. https://doi.org/10.1534/genetics.119.302299

LIANG, J. and SCHÖNLIEB, C.-B. (2018). Improving FISTA: Faster, smarter and greedier. ArXiv preprint. Available at arXiv:1811.01430.

LIANG, J. W. and SEN, Ś. (2022a). Supplemental figures for "Sparse matrix linear models for structured high-throughput data." https://doi.org/10.1214/21-AOAS1444SUPPA.

LIANG, J. W. and SEN, Ś. (2022b). matrixLMnet.jl Julia package for "Sparse matrix linear models for structured high-throughput data." https://doi.org/10.1214/21-AOAS1444SUPPB

LIANG, J. W. and SEN, Ś. (2022c). Code to reproduce analysis for "Sparse matrix linear models for structured high-throughput data." https://doi.org/10.1214/21-AOAS1444SUPPC

LOVELL, J. T., MULLEN, J. L., LOWRY, D. B., AWOLE, K., RICHARDS, J. H., SEN, S., VERSLUES, P. E., JUENGER, T. E. and MCKAY, J. K. (2015). Exploiting differential gene expression and epistasis to discover candidate genes for drought-associated QTLs in Arabidopsis thaliana. *Plant Cell* **27** 969–983.

LOWRY, D. B., LOGAN, T. L., SANTUARI, L., HARDTKE, C. S., RICHARDS, J. H., DEROSE-WILSON, L. J., MCKAY, J. K., SEN, S. and JUENGER, T. E. (2013). Expression quantitative trait locus mapping across water availability environments reveals contrasting associations with genomic features in Arabidopsis. *Plant Cell* **25** 3266–3279.

NESTEROV, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math., Dokl.* **27** 372–376. MR0701288

NICHOLS, R. J., SEN, S., CHOO, Y. J., BELTRAO, P., ZIETEK, M., CHABA, R., LEE, S., KAZMIERCZAK, K. M., LEE, K. J. et al. (2011). Phenotypic landscape of a bacterial cell. *Cell* **144** 143–156.

OCHS, P. and POCK, T. (2019). Adaptive FISTA for nonconvex optimization. *SIAM J. Optim.* **29** 2482–2503. MR4014792 https://doi.org/10.1137/17M1156678

PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 123–231.

R CORE TEAM (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2016). A study of error variance estimation in Lasso regression. *Statist. Sinica* **26** 35–67. MR3468344

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. and SMYTH, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** e47–e47.

SCHMIDT, M., FUNG, G. and ROSALES, R. (2009). Optimization methods for l1-regularization. University of British Columbia, Technical Report TR-2009 19.

SU, W., BOYD, S. and CANDÈS, E. J. (2016). A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.* **17** 153. MR3555044

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

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

TEAM, M. C., BLANCHARD, G., DICKHAUS, T., HACK, N., KONIETSCHKE, F., ROHMEYER, K., ROSENBLATT, J., SCHEER, M. and WERFT, W. (2017). mutoss: Unified Multiple Testing Procedures. R package version 0.1-12.

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

WOODRUFF, T. J., ZOTA, A. R. and SCHWARTZ, J. M. (2011). Environmental chemicals in pregnant women in the United States: NHANES 2003-2004. *Environ. Health Perspect.* **119** 878–885. https://doi.org/10.1289/ehp.1002727

WU, T. T. and LANGE, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601 https://doi.org/10.1214/07-AOAS147

XIONG, H., GOULDING, E. H., CARLSON, E. J., TECOTT, L. H., MCCULLOCH, C. E. and SEN, Ś. (2011). A flexible estimating equations approach for mapping function-valued traits. *Genetics* **189** 305–316.

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
2022, Vol. 16, No. 1, 193–215
https://doi.org/10.1214/21-AOAS1495

BIDIMENSIONAL LINKED MATRIX FACTORIZATION FOR PAN-OMICS PAN-CANCER ANALYSIS

BY ERIC F. LOCK[1,a], JUN YOUNG PARK[2,b] AND KATHERINE A. HOADLEY[3,c]

[1]*Division of Biostatistics, School of Public Health, University of Minnesota,* [a]*elock@umn.edu*
[2]*Department of Statistical Sciences, Faculty of Arts & Science, University of Toronto,* [b]*junjy.park@utoronto.ca*
[3]*Department of Genetics, Computational Medicine Program, University of North Carolina,* [c]*hoadley@med.unc.edu*

Several modern applications require the integration of multiple large data matrices that have shared rows and/or columns. For example, cancer studies that integrate multiple omics platforms across multiple types of cancer, *pan-omics pan-cancer analysis*, have extended our knowledge of molecular heterogeneity beyond what was observed in single tumor and single platform studies. However, these studies have been limited by available statistical methodology. We propose a flexible approach to the simultaneous factorization and decomposition of variation across such *bidimensionally linked* matrices, BIDIFAC+. BIDIFAC+ decomposes variation into a series of low-rank components that may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., cancer types). This builds on a growing literature for the factorization and decomposition of linked matrices which has primarily focused on multiple matrices that are linked in one dimension (rows or columns) only. Our objective function extends nuclear norm penalization, is motivated by random matrix theory, gives a unique decomposition under relatively mild conditions, and can be shown to give the mode of a Bayesian posterior distribution. We apply BIDIFAC+ to pan-omics pan-cancer data from TCGA, identifying shared and specific modes of variability across *four* different omics platforms and 29 different cancer types.

REFERENCES

AKBANI, R., NG, P. K. S., WERNER, H. M. J., SHAHMORADGOLI, M., ZHANG, F., JU, Z., LIU, W., YANG, J.-Y., YOSHIHARA, K. et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5** 3887. https://doi.org/10.1038/ncomms4887

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

GABASOVA, E., REID, J. and WERNISCH, L. (2017). Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput. Biol.* **13** e1005781. https://doi.org/10.1371/journal.pcbi.1005781

GAVISH, M. and DONOHO, D. L. (2017). Optimal shrinkage of singular values. *IEEE Trans. Inf. Theory* **63** 2137–2152. MR3626861 https://doi.org/10.1109/TIT.2017.2653801

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

HELLTON, K. H. and THORESEN, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics* **17** 537–548. MR3603952 https://doi.org/10.1093/biostatistics/kxw005

HOADLEY, K. A., YAU, C., WOLF, D. M., CHERNIACK, A. D., TAMBORERO, D., NG, S., LEISERSON, M. D. M., NIU, B., MCLELLAN, M. D. et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158** 929–944. https://doi.org/10.1016/j.cell.2014.06.049

HOADLEY, K. A., YAU, C., HINOUE, T., WOLF, D. M., LAZAR, A. J., DRILL, E., SHEN, R., TAYLOR, A. M., CHERNIACK, A. D. et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173** 291–304.

Key words and phrases. Cancer genomics, data integration, low-rank matrix factorization, missing data imputation, nuclear norm penalization.

HUO, Z. and TSENG, G. (2017). Integrative sparse K-means with overlapping group lasso in genomic applications for disease subtype discovery. *Ann. Appl. Stat.* **11** 1011–1039. MR3693556 https://doi.org/10.1214/17-AOAS1033

HUTTER, C. and ZENKLUSEN, J. C. (2018). The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell* **173** 283–285. https://doi.org/10.1016/j.cell.2018.03.042

KANDOTH, C., MCLELLAN, M. D., VANDIN, F., YE, K., NIU, B., LU, C., XIE, M., ZHANG, Q., MCMICHAEL, J. F. et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* **502** 333–339. https://doi.org/10.1038/nature12634

KAPLAN, A. and LOCK, E. F. (2017). Prediction with dimension reduction of multiple molecular data sources for patient survival. *Cancer Inform.* **16** 1–11.

KURUCZ, M., BENCZÚR, A. A. and CSALOGÁNY, K. (2007). Methods for large scale svd with missing values. In *Proceedings of KDD Cup and Workshop* **12** 31–38.

LI, G. and GAYNANOVA, I. (2018). A general framework for association analysis of heterogeneous data. *Ann. Appl. Stat.* **12** 1700–1726. MR3852694 https://doi.org/10.1214/17-AOAS1127

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

LOCK, E. F. and DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29** 2610–2616.

LOCK, E. F., PARK, J. Y. and HOADLEY, K. A. (2022). Supplement to "Bidimensional linked matrix factorization for pan-omics pan-cancer analysis." https://doi.org/10.1214/21-AOAS1495SUPPA, https://doi.org/10.1214/21-AOAS1495SUPPB, https://doi.org/10.1214/21-AOAS1495SUPPC

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

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

MO, Q., SHEN, R., GUO, C., VANNUCCI, M., CHAN, K. S. and HILSENBECK, S. G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19** 71–86. MR3799604 https://doi.org/10.1093/biostatistics/kxx017

NETWORK, T. R. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372** 2481–2498.

NETWORK, T. R. et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature* **490** 61.

NETWORK, T. R. et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511** 543.

NOUSHMEHR, H., WEISENBERGER, D. J., DIEFES, K., PHILLIPS, H. S., PUJARA, K., BERMAN, B. P., PAN, F., PELLOSKI, C. E., SULMAN, E. P. et al. (2010). Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17** 510–522.

O'CONNELL, M. J. and LOCK, E. F. (2016). R. JIVE for exploration of multi-source molecular data. *Bioinformatics* **32** 2877–2879.

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

PARK, J. Y. and LOCK, E. F. (2020). Integrative factorization of bidimensionally linked matrices. *Biometrics* **76** 61–74. MR4098544 https://doi.org/10.1111/biom.13141

RUDELSON, M. and VERSHYNIN, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III* 1576–1602. Hindustan Book Agency, New Delhi. MR2827856

SHABALIN, A. A. and NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.* **118** 67–76. MR3054091 https://doi.org/10.1016/j.jmva.2013.03.005

SHEN, R., WANG, S. and MO, Q. (2013). Sparse integrative clustering of multiple omics data sets. *Ann. Appl. Stat.* **7** 269–294. MR3086419 https://doi.org/10.1214/12-AOAS578

VERHAAK, R. G. W., HOADLEY, K. A., PURDOM, E., WANG, V., QI, Y., WILKERSON, M. D., MILLER, C. R., DING, L., GOLUB, T. et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17** 98–110. https://doi.org/10.1016/j.ccr.2009.12.020

WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C., STUART, J. M. et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45** 1113–1120.

YANG, Z. and MICHAILIDIS, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32** 1–8. https://doi.org/10.1093/bioinformatics/btv544

ZACK, T. I., SCHUMACHER, S. E., CARTER, S. L., CHERNIACK, A. D., SAKSENA, G., TABAK, B., LAWRENCE, M. S., ZHANG, C.-Z., WALA, J. et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45** 1134–1140.

ZHANG, S., WANG, W., FORD, J., MAKEDON, F. and PEARLMAN, J. (2005). Using singular value decomposition approximation for collaborative filtering. In *Seventh IEEE International Conference on E-Commerce Technology* (*CEC* '05) 257–264. IEEE.

ZHU, H., LI, G. and LOCK, E. F. (2020). Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *Biostatistics* **21** 302–318. MR4133362 https://doi.org/10.1093/biostatistics/kxy052

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 216–246
https://doi.org/10.1214/21-AOAS1477

A FUNCTIONAL-DATA APPROACH TO THE ARGO DATA

BY DREW YARGER[a], STILIAN STOEV[b] AND TAILEN HSING[c]

Department of Statistics, University of Michigan, [a]*dyarger@umich.edu,* [b]*sstoev@umich.edu,* [c]*thsing@umich.edu*

The Argo data is a modern oceanography dataset that provides unprecedented global coverage of temperature and salinity measurements in the upper 2000 meters of depth of the ocean. We study the Argo data from the perspective of functional data analysis (FDA). We develop spatiotemporal *functional kriging* methodology for mean and covariance estimation to predict temperature and salinity at a fixed location as a smooth function of depth. By combining tools from FDA and spatial statistics, including smoothing splines, local regression, and multivariate spatial modeling and prediction, our approach provides advantages over current methodology that consider pointwise estimation at fixed depths. Our approach naturally leverages the irregularly-sampled data in space, time, and depth to fit a space-time functional model for temperature and salinity. The developed framework provides new tools to address fundamental scientific problems involving the entire upper water column of the oceans, such as the estimation of ocean heat content, stratification, and thermohaline oscillation. For example, we show that our functional approach yields more accurate ocean heat content estimates than ones based on discrete integral approximations in pressure. Further, using the derivative function estimates, we obtain a new product of a global map of the mixed layer depth, a key component in the study of heat absorption and nutrient circulation in the oceans. The derivative estimates also reveal evidence for density inversions in areas distinguished by mixing of particularly different water masses.

REFERENCES

AGUILERA-MORILLO, M. C., DURBÁN, M. and AGUILERA, A. M. (2017). Prediction of functional data with spatial dependence: A penalized approach. *Stoch. Environ. Res. Risk Assess.* **31** 7–22. https://doi.org/10.1007/s00477-016-1216-8

ARGO (2000). Argo float data and metadata from global data assembly centre (Argo GDAC). *SEANOE*.

BACHOC, F., GENTON, M. G., NORDHAUSEN, K., RUIZ-GAZEN, A. and VIRTA, J. (2020). Spatial blind source separation. *Biometrika* **107** 627–646. MR4138980 https://doi.org/10.1093/biomet/asz079

BALADANDAYUTHAPANI, V., MALLICK, B. K., HONG, M. Y., LUPTON, J. R., TURNER, N. D. and CARROLL, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64** 64–73, 321–322. MR2422820 https://doi.org/10.1111/j.1541-0420.2007.00846.x

BARTH, A., AZCARATE, A. A., JOASSIN, P., BECKERS, J.-M. and TROUPIN, C. (2008). Introduction to optimal interpolation and variational analysis. *GeoHydrodynamics and Environmental Research* 27.

BOLIN, D. and WALLIN, J. (2020). Multivariate type G Matérn stochastic partial differential equation random fields. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 215–239. MR4060983

BOYER, T. P. and LEVITUS, S. (1994). *Quality Control and Processing of Historical Oceanographic Temperature, Salinity, and Oxygen Data* **81**. US Department of Commerce, National Oceanic and Atmospheric Administration.

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** 1190–1208. MR1346301 https://doi.org/10.1137/0916069

CHENG, L. and ZHU, J. (2014). Uncertainties of the ocean heat content estimation induced by insufficient vertical resolution of historical ocean subsurface observations. *J. Atmos. Ocean. Technol.* **31** 1383–1396. https://doi.org/10.1175/JTECH-D-13-00220.1

CHILÈS, J.-P. and DELFINER, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2850475 https://doi.org/10.1002/9781118136188

Key words and phrases. Functional data analysis, Matérn, oceanography, spatial statistics, splines.

CHOI, H. and REIMHERR, M. (2018). A geometric approach to confidence regions and bands for functional parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 239–260. MR3744720 https://doi.org/10.1111/rssb.12239
CRESSIE, N. and WIKLE, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons. MR2848400
CRONIE, O., GHORBANI, M., MATEU, J. and YU, J. (2019). Functional marked point processes—a natural structure to unify spatio-temporal frameworks and to analyse dependent functional data. arXiv:1911.13142.
DELICADO, P., GIRALDO, R., COMAS, C. and MATEU, J. (2010). Statistics for spatial functional data: Some recent contributions. *Environmetrics* **21** 224–239. MR2842240 https://doi.org/10.1002/env.1003
FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587
GAILLARD, F. (2012). ISAS-tool version 6: Method and configuration.
GRAY, A. R. and RISER, S. C. (2015). A method for multiscale optimal analysis with application to Argo data. *J. Geophys. Res.* **120** 4340–4356. https://doi.org/10.1002/2014JC010208
GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*: *A Roughness Penalty Approach*. *Monographs on Statistics and Applied Probability* **58**. CRC Press, London. MR1270012 https://doi.org/10.1007/978-1-4899-4473-3
GROMENKO, O., KOKOSZKA, P. and SOJKA, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *Ann. Appl. Stat.* **11** 898–918. MR3693551 https://doi.org/10.1214/17-AOAS1022
GUINNESS, J. (2021). Gaussian process learning via Fisher scoring of Vecchia's approximation. *Stat. Comput.* **31** Paper No. 25, 8. MR4224950 https://doi.org/10.1007/s11222-021-09999-1
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
HOLTE, J. and TALLEY, L. (2009). A new algorithm for finding mixed layer depths with applications to Argo data and Subantarctic mode water formation. *J. Atmos. Ocean. Technol.* **26** 1920–1939.
HOLTE, J., TALLEY, L. D., GILSON, J. and ROEMMICH, D. (2017). An Argo mixed layer climatology and database. *Geophys. Res. Lett.* **44** 5618–5626. https://doi.org/10.1002/2017GL073426
HOSODA, S., OHIRA, T. and NAKAMURA, T. (2008). A monthly mean dataset of global oceanic temperature and salinity derived from Argo float observations. *JAMSTEC Report of Research and Development* **8** 47–59. https://doi.org/10.5918/jamstecr.8.47
HOSODA, S., OHIRA, T., SATO, K. and SUGA, T. (2010). Improved description of global mixed-layer depth using Argo profiling floats. *J. Oceanogr.* **66** 773–787.
HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547
ISHII, M. and KIMOTO, M. (2009). Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanogr.* **65** 287–299.
JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427 https://doi.org/10.1080/00401706.2012.657106
JOHNSON, G. C. and BIRNBAUM, A. N. (2017). As El Niño builds, Pacific warm pool expands, ocean gains more heat. *Geophys. Res. Lett.* **44** 438–445. https://doi.org/10.1002/2016GL071767
KELLEY, D., RICHARDS, C. and WG127 SCOR/IAPSO (2017). gsw: Gibbs sea water functions. R package version 1.0-5.
KING, M. C., STAICU, A.-M., DAVIS, J. M., REICH, B. J. and EDER, B. (2018). A functional data analysis of spatiotemporal trends and variation in fine particulate matter. *Atmos. Environ.* **184** 233–243. https://doi.org/10.1016/j.atmosenv.2018.04.001
KOKOSZKA, P. and REIMHERR, M. (2017). *Introduction to Functional Data Analysis*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3793167
KOKOSZKA, P. and REIMHERR, M. (2019). Some recent developments in inference for geostatistical functional data. *Rev. Colombiana Estadíst.* **42** 101–122. MR3919840 https://doi.org/10.15446/rce.v42n1.77058
KUUSELA, M. and STEIN, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **474**.
LEVITUS, S., ANTONOV, J. I., BOYER, T. P., BARANOVA, O. K., GARCIA, H. E., LOCARNINI, R. A., MISHONOV, A. V., REAGAN, J. R., SEIDOV, D. et al. (2012). World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophys. Res. Lett.* **39**. https://doi.org/10.1029/2012GL051106
LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. MR2766854 https://doi.org/10.1214/10-AOS813
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

LI, H., XU, F., ZHOU, W., WANG, D., WRIGHT, J., LIU, Z. and LIN, Y. (2017). Development of a global gridded Argo data set with Barnes successive corrections. *J. Geophys. Res.* **122**. https://doi.org/10.1002/2016JC012285
LI, G., CHENG, L., ZHU, J., TRENBERTH, K. E., MANN, M. E. and ABRAHAM, J. P. (2020). Increasing ocean stratification over the past half-century. *Nat. Clim. Change* 1–8. https://doi.org/10.1038/s41558-020-00918-2
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
LYMAN, J. M. and JOHNSON, G. C. (2013). Estimating global ocean heat content changes in the upper 1800 m since 1950 and the influence of climatology choice. *J. Climate* **27** 1945–1957. https://doi.org/10.1175/JCLI-D-12-00752.1
LYNCH, B. and CHEN, K. (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika* **105** 815–831. MR3877867 https://doi.org/10.1093/biomet/asy048
MARTÍNEZ-HERNÁNDEZ, I. and GENTON, M. G. (2020). Recent developments in complex and spatially correlated functional data. *Braz. J. Probab. Stat.* **34** 204–229. MR4093256 https://doi.org/10.1214/20-BJPS466
MCDOUGALL, T. J. (2003). Potential enthalpy: A conservative oceanic variable for evaluating heat content and heat fluxes. *J. Phys. Oceanogr.* **33** 945–963. MR2001123 https://doi.org/10.1175/1520-0485(2003)033<0945:PEACOV>2.0.CO;2
MCDOUGALL, T. J. and BARKER, P. M. (2011). Getting started with TEOS-10 and the Gibbs seawater (GSW) oceanographic toolbox. *SCOR/IAPSO WG*127 **33** 28.
MEYSSIGNAC, B., BOYER, T., ZHAO, Z., HAKUBA, M. Z., LANDERER, F. W., STAMMER, D., KÖHL, A., KATO, S., L'ECUYER, T. et al. (2019). Measuring global ocean heat content to estimate the Earth energy imbalance. *Front. Mar. Sci.* **6** 1–31. https://doi.org/10.3389/fmars.2019.00432
MONESTIEZ, P. and NERINI, D. (2008). A cokriging method for spatial functional data with applications in oceanology. In *Functional and Operatorial Statistics. Contrib. Statist.* 237–242. Physica-Verlag/Springer, Heidelberg. MR2490355 https://doi.org/10.1007/978-3-7908-2062-1_36
NOAA NODC (2019). Ocean climate laboratory, global ocean heat and salt content global anomaly fields. *National Oceanographic Data Center*.
OWENS, W. B. and WONG, A. P. S. (2009). An improved calibration method for the drift of the conductivity sensor on autonomous CTD profiling floats by θ–S climatology. *Deep-Sea Res., Part* 1, *Oceanogr. Res. Pap.* **56** 450–457. https://doi.org/10.1016/j.dsr.2008.09.008
PAUTHENET, E., ROQUET, F., MADEC, G., SALLÉE, J.-B. and NERINI, D. (2019). The thermohaline modes of the global ocean. *J. Phys. Oceanogr.* **49** 2535–2552. https://doi.org/10.1175/JPO-D-19-0120.1
RAHMSTORF, S., BOX, J. E., FEULNER, G., MANN, M. E., ROBINSON, A., RUTHERFORD, S. and SCHAFFERNICHT, E. J. (2015). Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Change* **5** 475–480. https://doi.org/10.1038/nclimate2554
RAMSAY, J. and SILVERMAN, B. W. (2013). *Functional Data Analysis. Springer Series in Statistics*. Springer Science & Business Media, New York. MR2168993
RAMSAY, J., WICKHAM, H., GRAVES, S. and HOOKER, G. (2018). fda: Functional data analysis. R package version 2.4.8.
RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96** 149–162. MR2482141 https://doi.org/10.1093/biomet/asn054
ROEMMICH, D. and GILSON, J. (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo program. *Prog. Oceanogr.* **82** 81–100.
ROEMMICH, D., GOULD, W. J. and GILSON, J. (2012). 135 years of global ocean warming between the Challenger expedition and the Argo programme. *Nat. Clim. Change* **2** 425–428. https://doi.org/10.1038/nclimate1461
ROEMMICH, D., CHURCH, J., GILSON, J., MONSELESAN, D., SUTTON, P. and WIJFFELS, S. (2015). Unabated planetary warming and its ocean structure since 2006. *Nat. Clim. Change* **5** 240–245. https://doi.org/10.1038/nclimate2513
ROMANO, E., BALZANELLA, A. and VERDE, R. (2017). Spatial variability clustering for spatially dependent functional data. *Stat. Comput.* **27** 645–658. MR3613590 https://doi.org/10.1007/s11222-016-9645-2
RUIZ-MEDINA, M. D. (2011). Spatial autoregressive and moving average Hilbertian processes. *J. Multivariate Anal.* **102** 292–305. MR2739116 https://doi.org/10.1016/j.jmva.2010.09.005
RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comput. Graph. Statist.* **11** 735–757. MR1944261 https://doi.org/10.1198/106186002321018768
SCHMIDTKO, S., JOHNSON, G. C. and LYMAN, J. M. (2013). MIMOC: A global monthly isopycnal upper-ocean climatology with mixed layers. *J. Geophys. Res.* **118** 1658–1672.
SONG, J. J. and MALLICK, B. (2019). Hierarchical Bayesian models for predicting spatially correlated curves. *Statistics* **53** 196–209. MR3900086 https://doi.org/10.1080/02331888.2018.1547905

STAICU, A.-M., CRAINICEANU, C. M. and CARROLL, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostat*. **11** 177–194. https://doi.org/10.1093/biostatistics/kxp058

STAICU, A.-M., CRAINICEANU, C. M., REICH, D. S. and RUPPERT, D. (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics* **68** 331–343. MR2959599 https://doi.org/10.1111/j.1541-0420.2011.01669.x

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. *Springer Series in Statistics*. Springer, New York. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

TALLEY, L. D., PICKARD, G. L., EMERY, W. J. and SWIFT, J. H. (2011). *Descriptive Physical Oceanography*, 6th ed. Academic Press, Boston, MA. https://doi.org/10.1016/B978-0-7506-4552-2.10004-6

THOMSON, R. E. and EMERY, W. J. (2014). *Data Analysis Methods in Physical Oceanography*, 3rd edition. Elsevier, Boston, MA. https://doi.org/10.1016/B978-0-12-387782-6.00003-X

TOWNS, J., COCKERILL, T., DAHAN, M., FOSTER, I., GAITHER, K., GRIMSHAW, A., HAZLEWOOD, V., LATHROP, S., LIFKA, D. et al. (2014). XSEDE: Accelerating scientific discovery. *Comput*. *Sci*. *Eng*. **16** 62–74. https://doi.org/10.1109/MCSE.2014.80

UDAYA BHASKAR, T. V. S., RAVICHANDRAN, M. and DEVENDER, R. (2007). An operational objective analysis system at INCOIS for generation of Argo value added products.

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

WIENS, A., NYCHKA, D. and KLEIBER, W. (2020). Modeling spatial data using local likelihood estimation and a Matérn to spatial autoregressive translation. *Environmetrics* **31** A1E2652–14. MR4151871 https://doi.org/10.1002/env.2652

WOOD, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62** 1025–1036. MR2297673 https://doi.org/10.1111/j.1541-0420.2006.00574.x

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

YARGER, D. (2020a). Shiny applications accompanying "A functional-data approach to the Argo data". Available at https://sites.google.com/a/umich.edu/argostatistics/home/fdapaper.

YARGER, D. (2020b). Code accompanying "A functional-data approach to the Argo data". Available at https://github.com/dyarger/argofda.

YARGER, D., STOEV, S. and HSING, T. (2022). Supplement to "A functional-data approach to the Argo data." https://doi.org/10.1214/21-AOAS1477SUPPA, https://doi.org/10.1214/21-AOAS1477SUPPB

ZHANG, H. and LI, Y. (2020). Unified principal component analysis for sparse and dense functional data under spatial dependency. arXiv:2006.13489.

ZHANG, L., BALADANDAYUTHAPANI, V., ZHU, H., BAGGERLY, K. A., MAJEWSKI, T., CZERNIAK, B. A. and MORRIS, J. S. (2016). Functional CAR models for large spatially correlated functional datasets. *J*. *Amer*. *Statist*. *Assoc*. **111** 772–786. MR3538704 https://doi.org/10.1080/01621459.2015.1042581

ZHOU, L., HUANG, J. Z., MARTINEZ, J. G., MAITY, A., BALADANDAYUTHAPANI, V. and CARROLL, R. J. (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *J*. *Amer*. *Statist*. *Assoc*. **105** 390–400. MR2757206 https://doi.org/10.1198/jasa.2010.tm08737

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 247–271
https://doi.org/10.1214/21-AOAS1497

FAST INFERENCE FOR TIME-VARYING QUANTILES VIA FLEXIBLE DYNAMIC MODELS WITH APPLICATION TO THE CHARACTERIZATION OF ATMOSPHERIC RIVERS

BY RAQUEL BARATA[a], RAQUEL PRADO[b] AND BRUNO SANSÓ[c]

Department of Statistics, University of California Santa Cruz, [a]*rbarata@ucsc.udu,* [b]*raquel@soe.ucsc.edu,* [c]*bruno@soe.ucsc.edu*

Atmospheric rivers (ARs) are elongated regions of water vapor in the atmosphere that play a key role in global water cycles, particularly in western U.S. precipitation. The primary component of many AR detection schemes is the thresholding of the integrated water vapor transport (IVT) magnitude at a single quantile over time. Utilizing a recently developed family of parametric distributions for quantile regression, this paper develops a flexible dynamic quantile linear model (exDQLM) which enables versatile, structured, and informative estimation of the IVT quantile threshold. A simulation study illustrates our exDQLM to be more robust than the standard Bayesian parametric quantile regression approach for nonstandard distributions, performing better in both quantile estimation and predictive accuracy. In addition to a Markov chain Monte Carlo (MCMC) algorithm, we develop an efficient importance sampling variational Bayes (ISVB) algorithm for fast approximate Bayesian inference which is found to produce comparable results to the MCMC in a fraction of the computation time. Further, we develop a transfer function extension to our exDQLM as a method for quantifying nonlinear relationships between a quantile of a climatological response and an input. The utility of our transfer function exDQLM is demonstrated in capturing both the immediate and lagged effects of El Niño Southern Oscillation Longitude Index on the estimation of the 0.85 quantile IVT.

REFERENCES

BACKES, T. M., KAPLAN, M. L., SCHUMER, R. and MEJIA, J. F. (2015). A climatology of the vertical structure of water vapor transport to the Sierra Nevada in cool season atmospheric river precipitation events. *J. Hydrometeorol.* **16** 1029–1047.

BARATA, R., PRADO, R. and SANSÓ, B. (2022). Supplement to "Fast inference for time-varying quantiles via flexible dynamic models with application to the characterization of atmospheric rivers." https://doi.org/10.1214/21-AOAS1497SUPP

BARBER, D. and CHIAPPA, S. (2007). Unified inference for variational Bayesian linear Gaussian state-space models. In *Advances in Neural Information Processing Systems* 81–88.

BEAL, M. J. (2003). Variational algorithms for approximate Bayesian inference. Ph.D. thesis, UCL (Univ. College London).

BERNARDI, M., CASARIN, R., MAILLET, B. and PETRELLA, L. (2016). Dynamic model averaging for Bayesian quantile regression. Preprint. Available at arXiv:1602.00856.

BERRISFORD, P., KÅLLBERG, P., KOBAYASHI, S., DEE, D., UPPALA, S., SIMMONS, A., POLI, P. and SATO, H. (2011). Atmospheric conservation properties in ERA-interim. *Q. J. R. Meteorol. Soc.* **137** 1381–1399.

BERRY, L. R. and WEST, M. (2020). Bayesian forecasting of many count-valued time series. *J. Bus. Econom. Statist.* **38** 872–887. MR4154894 https://doi.org/10.1080/07350015.2019.1604372

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

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096 https://doi.org/10.1093/biomet/81.3.541

CHEN, W. Y., PETERS, G. W., GERLACH, R. H. and SISSON, S. A. (2017). Dynamic quantile function models. Preprint. Available at arXiv:1707.02587.

Key words and phrases. Dynamic quantile regression, asymmetric Laplace, variational Bayes, atmospheric river.

DEE, D. P., UPPALA, S., SIMMONS, A., BERRISFORD, P., POLI, P., KOBAYASHI, S., ANDRAE, U., BALMASEDA, M., BALSAMO, G. et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137** 553–597.

FOTI, N. J., XU, J., LAIRD, D. and FOX, E. B. (2014). Stochastic variational inference for hidden Markov models. Preprint. Available at arXiv:1411.1670.

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889 https://doi.org/10.1111/j.1467-9892.1994.tb00184.x

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11. MR1627258 https://doi.org/10.1093/biomet/85.1.1

GONÇALVES, K. C. M., MIGON, H. S. and BASTOS, L. S. (2020). Dynamic quantile linear models: A Bayesian approach. *Bayesian Anal.* **15** 335–362. MR4078717 https://doi.org/10.1214/19-BA1156

GUAN, B. and WALISER, D. E. (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res., Atmos.* **120** 12514–12535.

GUAN, B., WALISER, D. E., MOLOTCH, N. P., FETZER, E. J. and NEIMAN, P. J. (2012). Does the Madden–Julian oscillation influence wintertime atmospheric rivers and snowpack in the Sierra Nevada? *Mon. Weather Rev.* **140** 325–342.

GUAN, B., MOLOTCH, N. P., WALISER, D. E., FETZER, E. J. and NEIMAN, P. J. (2013). The 2010/2011 snow season in California's Sierra Nevada: Role of atmospheric rivers and modes of large-scale variability. *Water Resour. Res.* **49** 6731–6743.

HANSON, T. and JOHNSON, W. O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97** 1020–1033. MR1951256 https://doi.org/10.1198/016214502388618843

HENZE, N. (1986). A probabilistic representation of the "skew-normal" distribution. *Scand. J. Stat.* **13** 271–275. MR0886466

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.

HOSSZEJNI, D. and KASTNER, G. (2018). Approaches toward the Bayesian estimation of the stochastic volatility model with leverage. In *International Conference on Bayesian Statistics in Action* 75–83. Springer, Berlin.

HUERTA, G., JIANG, W. and TANNER, M. A. (2003). Time series modeling via hierarchical mixtures. *Statist. Sinica* 1097–1118.

JOHNSON, M. and WILLSKY, A. (2014). Stochastic variational inference for Bayesian time series models. In *International Conference on Machine Learning* 1854–1862. PMLR.

KASTNER, G. (2016). Dealing with stochastic volatility in time series using the R package stochvol. *J. Stat. Softw.* **69** 1–30.

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657 https://doi.org/10.1017/CBO9780511754098

KOENKER, R. and XIAO, Z. (2006). Quantile autoregression. *J. Amer. Statist. Assoc.* **101** 980–990. MR2324109 https://doi.org/10.1198/016214506000000672

KOMUNJER, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *J. Econometrics* **128** 137–164. MR2022929 https://doi.org/10.1016/j.jeconom.2004.08.010

KOTTAS, A. and GELFAND, A. E. (2001). Bayesian semiparametric mdeian regression modeling. *J. Amer. Statist. Assoc.* **96** 1458–1468. MR1946590 https://doi.org/10.1198/016214501753382363

KOTTAS, A. and KRNJAJIĆ, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scand. J. Stat.* **36** 297–319. MR2528986 https://doi.org/10.1111/j.1467-9469.2008.00626.x

KOTZ, S., KOZUBOWSKI, T. J. and PODGÓRSKI, K. (2001). *The Laplace Distribution and Generalizations*: *A Revisit With Applications to Communications*, *Economics*, *Engineering*, *and Finance*. Birkhäuser, Boston, MA. MR1935481 https://doi.org/10.1007/978-1-4612-0173-1

KOZUMI, H. and KOBAYASHI, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81** 1565–1578. MR2851270 https://doi.org/10.1080/00949655.2010.496117

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

LISEO, B. and LOPERFIDO, N. (2006). A note on reference priors for the scalar skew-normal distribution. *J. Statist. Plann. Inference* **136** 373–389. MR2211345 https://doi.org/10.1016/j.jspi.2004.06.062

LUM, K. and GELFAND, A. E. (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Anal.* **7** 235–258. MR2934947 https://doi.org/10.1214/12-BA708

NEELON, B., LI, F., BURGETTE, L. F. and BENJAMIN NEELON, S. E. (2015). A spatiotemporal quantile regression model for emergency department expenditures. *Stat. Med.* **34** 2559–2575. MR3368401 https://doi.org/10.1002/sim.6480

NEIMAN, P. J., WHITE, A. B., RALPH, F. M., GOTTAS, D. J. and GUTMAN, S. I. (2009). A water vapour flux tool for precipitation forecasting. In *Proceedings of the Institution of Civil Engineers-Water Management* **162** 83–94. Thomas Telford Ltd.

OSTWALD, D., KIRILINA, E., STARKE, L. and BLANKENBURG, F. (2014). A tutorial on variational Bayes for latent linear stochastic time-series models. *J. Math. Psych.* **60** 1–19. MR3245721 https://doi.org/10.1016/j.jmp.2014.04.003

PARASCHIV, F., BUNN, D. and WESTGAARD, S. (2016). Estimation and application of fully parametric multifactor quantile regression with dynamic coefficients. Univ. St. Gallen, School of Finance Research Paper 2016/07.

PATRICOLA, C. M., O'BRIEN, J. P., RISSER, M. D., RHOADES, A. M., O'BRIEN, T. A., ULLRICH, P. A., STONE, D. A. and COLLINS, W. D. (2020). Maximizing ENSO as a source of western US hydroclimate predictability. *Clim. Dyn.* **54** 351–372.

PENNY, W., KIEBEL, S. and FRISTON, K. (2003). Variational Bayesian inference for fMRI time series. *NeuroImage* **19** 727–741.

PRADO, R., MOLINA, F. and HUERTA, G. (2006). Multivariate time series modeling and classification via hierarchical VAR mixtures. *Comput. Statist. Data Anal.* **51** 1445–1462. MR2307518 https://doi.org/10.1016/j.csda.2006.03.002

QUIROZ, M., NOTT, D. J. and KOHN, R. (2018). Gaussian variational approximation for high-dimensional state space models. Preprint. Available at arXiv:1801.07873.

REICH, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 535–553. MR2960737 https://doi.org/10.1111/j.1467-9876.2011.01025.x

REICH, B. J., BONDELL, H. D. and WANG, H. J. (2009). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* **11** 337–352.

REICH, B. J., FUENTES, M. and DUNSON, D. B. (2011). Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.* **106** 6–20. MR2816698 https://doi.org/10.1198/jasa.2010.ap09237

ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525 https://doi.org/10.1214/aoms/1177729394

RUTZ, J. J., STEENBURGH, W. J. and RALPH, F. M. (2014). Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Weather Rev.* **142** 905–921.

TADDY, M. A. and KOTTAS, A. (2010). A Bayesian nonparametric approach to inference for quantile regression. *J. Bus. Econom. Statist.* **28** 357–369. MR2723605 https://doi.org/10.1198/jbes.2009.07331

TSIONAS, E. G. (2003). Bayesian quantile inference. *J. Stat. Comput. Simul.* **73** 659–674. MR2001612 https://doi.org/10.1080/0094965031000064463

TUCKERMAN, M. E. (2010). *Statistical Mechanics*: *Theory and Molecular Simulation. Oxford Graduate Texts*. Oxford Univ. Press, Oxford. MR2723222

TZIPERMAN, E., CANE, M. A., ZEBIAK, S. E., XUE, Y. and BLUMENTHAL, B. (1998). Locking of El Nino's peak time to the end of the calendar year in the delayed oscillator picture of ENSO. *J. Climate* **11** 2191–2199.

WALKER, S. and MALLICK, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics* **55** 477–483. MR1705102 https://doi.org/10.1111/j.0006-341X.1999.00477.x

WELLER, G. B., COOLEY, D. S. and SAIN, S. R. (2012). An investigation of the pineapple express phenomenon via bivariate extreme value theory. *Environmetrics* **23** 420–439. MR2958922 https://doi.org/10.1002/env.2143

WEST, M. and HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models. Springer Series in Statistics*. Springer, New York. MR1020301 https://doi.org/10.1007/978-1-4757-9365-9

WICHITAKSORN, N., CHOY, S. T. B. and GERLACH, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canad. J. Statist.* **42** 579–596. MR3281462 https://doi.org/10.1002/cjs.11228

WILLIAMS, I. N. and PATRICOLA, C. M. (2018). Diversity of ENSO events unified by convective threshold sea surface temperature: A nonlinear ENSO index. *Geophys. Res. Lett.* **45** 9236–9244.

YAN, Y. and KOTTAS, A. (2017). A new family of error distributions for Bayesian quantile regression. Preprint. Available at arXiv:1701.05666.

YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. MR1861390 https://doi.org/10.1016/S0167-7152(01)00124-9

ZHU, D. and GALBRAITH, J. W. (2011). Modeling and forecasting expected shortfall with the generalized asymmetric Student-t and asymmetric exponential power distributions. *J. Empir. Finance* **18** 765–778.

ZHU, D. and ZINDE-WALSH, V. (2009). Properties and estimation of asymmetric exponential power distribution. *J. Econometrics* **148** 86–99. MR2494820 https://doi.org/10.1016/j.jeconom.2008.09.038

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 272–299
https://doi.org/10.1214/21-AOAS1504

MODELING NONSTATIONARY TEMPERATURE MAXIMA BASED ON EXTREMAL DEPENDENCE CHANGING WITH EVENT MAGNITUDE

BY PENG ZHONG[1,a], RAPHAËL HUSER[1,??] AND THOMAS OPITZ[2,c]

[1]*Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST),* [a]*peng.zhong@kaust.edu.sa,* [b]*raphael.huser@kaust.edu.sa*
[2]*BioSP, INRAE,* [c]*thomas.opitz@inrae.fr*

The modeling of spatiotemporal trends in temperature extremes can help better understand the structure and frequency of heatwaves in a changing climate and assess the environmental, societal, economic and health-related risks they entail. Here, we study annual temperature maxima over Southern Europe using a century-spanning dataset observed at 44 monitoring stations. Extending the spectral representation of max-stable processes, our modeling framework relies on a novel construction of max-infinitely divisible processes which include covariates to capture spatiotemporal nonstationarities. Our new model keeps a popular max-stable process on the boundary of the parameter space, while flexibly capturing weakening extremal dependence at increasing quantile levels and asymptotic independence. This is achieved by linking the overall magnitude of a spatial event to its spatial correlation range in such a way that more extreme events become less spatially dependent, thus more localized. Our model reveals salient features of the spatiotemporal variability of European temperature extremes, and it clearly outperforms natural alternative models. Results show that the spatial extent of heatwaves is smaller for more severe events at higher elevations and that recent heatwaves are moderately wider. Our probabilistic assessment of the 2019 annual maxima confirms the severity of the 2019 heatwaves both spatially and at individual sites, especially when compared to climatic conditions prevailing in 1950–1975. Our results could be exploited in practice to understand the spatiotemporal dynamics, severity and frequency of extreme heatwaves and to design suitable region-specific mitigation measures.

REFERENCES

BALKEMA, A. A., DE HAAN, L. and KARANDIKAR, R. L. (1993). Asymptotic distribution of the maximum of n independent stochastic processes. *J. Appl. Probab.* **30** 66–81. MR1206353 https://doi.org/10.1017/s0021900200044004

BALLESTER, F., CORELLA, D., PÉREZ-HOYOS, S., SÁEZ, M. and HERVÁS, A. (1997). Mortality as a function of temperature. A study in Valencia, Spain, 1991–1993. *Int. J. Epidemiol.* **26** 551–561.

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. MR4227677 https://doi.org/10.1080/01621459.2020.1750414

BROWN, A. (2016). Heatwave mortality. *Nat. Clim. Change* **6** 821.

BROWN, B. M. and RESNICK, S. I. (1977). Extreme values of independent stochastic processes. *J. Appl. Probab.* **14** 732–739. MR0517438 https://doi.org/10.2307/3213346

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. MR3572037 https://doi.org/10.1080/10618600.2015.1086656

DAVISON, A. C. and GHOLAMREZAEE, M. M. (2012). Geostatistics of extremes. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **468** 581–608. MR2874052 https://doi.org/10.1098/rspa.2011.0412

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

Key words and phrases. Asymptotic dependence and independence, max-infinitely divisible process, max-stable process, spatial extreme event, spatiotemporal modeling.

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235.
DAVISON, A., HUSER, R. and THIBAUD, E. (2019). Spatial extremes. In *Handbook of Environmental and Ecological Statistics*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 711–744. CRC Press, Boca Raton, FL. MR3889918
DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980 https://doi.org/10.1214/11-STS376
DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. Roy. Statist. Soc. Ser. B* **52** 393–442. MR1086795
DE HAAN, L. (1984). A spectral representation for max-stable processes. *Ann. Probab.* **12** 1194–1204. MR0757776
DOMBRY, C. and EYI-MINKO, F. (2013). Regular conditional distributions of continuous max-infinitely divisible random fields. *Electron. J. Probab.* **18** no. 7, 21 pp. MR3024101 https://doi.org/10.1214/EJP.v18-1991
EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
FIELD, C. B., BARROS, V., STOCKER, T. F., QIN, D., DOKKEN, D. J., EBI, K. L., MASTRANDREA, M. D., MACH, K. J., PLATTNER, G. K. et al. (2012). *IPCC, 2012: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge Univ. Press, Cambridge, UK.
FISHER, R. A. and TIPPETT, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Philos. Soc.* **24** 180.
GENEST, C., GHOUDI, K. and RIVEST, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82** 543–552. MR1366280 https://doi.org/10.1093/biomet/82.3.543
GINÉ, E., HAHN, M. G. and VATAN, P. (1990). Max-infinitely divisible and max-stable sample continuous processes. *Probab. Theory Related Fields* **87** 139–165. MR1080487 https://doi.org/10.1007/BF01198427
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
GORDO, O. and SANZ, J. J. (2010). Impact of climate change on plant phenology in Mediterranean ecosystems. *Glob. Change Biol.* **16** 1082–1106.
GRABHERR, G., GOTTFRIED, M. and PAULI, H. (2010). Climate change impacts in alpine environments. *Geogr. Compass* **4** 1133–1153.
HUSER, R. and DAVISON, A. C. (2013). Composite likelihood estimation for the Brown–Resnick process. *Biometrika* **100** 511–518. MR3068451 https://doi.org/10.1093/biomet/ass089
HUSER, R. and DAVISON, A. C. (2014). Space-time modelling of extreme events. *J. Roy. Statist. Soc. Ser. B* **76** 439–461. MR3164873 https://doi.org/10.1111/rssb.12035
HUSER, R., DAVISON, A. C. and GENTON, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes* **19** 79–103. MR3454032 https://doi.org/10.1007/s10687-015-0230-4
HUSER, R. and GENTON, M. G. (2016). Non-stationary dependence structures for spatial extremes. *J. Agric. Biol. Environ. Stat.* **21** 470–491. MR3542082 https://doi.org/10.1007/s13253-016-0247-4
HUSER, R., OPITZ, T. and THIBAUD, E. (2017). Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spat. Stat.* **21** 166–186. MR3692183 https://doi.org/10.1016/j.spasta.2017.06.004
HUSER, R., OPITZ, T. and THIBAUD, E. (2021). Max-infinitely divisible models and inference for spatial extremes. *Scand. J. Stat.* **48** 321–348. MR4233175 https://doi.org/10.1111/sjos.12491
HUSER, R. and WADSWORTH, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *J. Amer. Statist. Assoc.* **114** 434–444. MR3941266 https://doi.org/10.1080/01621459.2017.1411813
HUSER, R. and WADSWORTH, J. L. (2020). Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev.* (*WIREs*): *Comput. Stat.* **e1537**. To appear.
HUSER, R., DOMBRY, C., RIBATET, M. and GENTON, M. G. (2019). Full likelihood inference for max-stable data. *Stat* **8** e218, 14 pp. MR3938281 https://doi.org/10.1002/sta4.218
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
JOE, H. (2015). *Dependence Modeling with Copulas. Monographs on Statistics and Applied Probability* **134**. CRC Press, Boca Raton, FL. MR3328438
JOE, H. and XU, J. J. (1996). The estimation method of inference functions for margins for multivariate models. Technical Report #166, Univ. British Columbia, Vancouver, Canada.
JUN, M., KNUTTI, R. and NYCHKA, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Statist. Assoc.* **103** 934–947. MR2528820 https://doi.org/10.1198/016214507000001265
KABLUCHKO, Z., SCHLATHER, M. and DE HAAN, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Probab.* **37** 2042–2065. MR2561440 https://doi.org/10.1214/09-AOP455

KLAUSMEYER, K. R. and SHAW, M. R. (2009). Climate change, habitat loss, protected areas and the climate adaptation potential of species in Mediterranean ecosystems worldwide. *PLoS ONE* **4** e6392. https://doi.org/10.1371/journal.pone.0006392

KLEIN TANK, A. M. G., WIJNGAARD, J. B., KÖNNEN, G. P., BÖHM, R., DEMARÉE, G., GOCHEVA, A., MILETA, M., PASHIARDIS, S., HEJKRLIK, L. et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.* **22** 1441–1453.

LEDFORD, A. W. and TAWN, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83** 169–187. MR1399163 https://doi.org/10.1093/biomet/83.1.169

LYON, B., BARNSTON, G. A., COFFEL, E. and HORTON, R. M. (2019). Projected increase in the spatial extent of contiguous US summer heat waves and associated attributes. *Environ. Res. Lett.* **14** 114029.

MITCHELL, D., HEAVISIDE, C., VARDOULAKIS, S., HUNTINGFORD, C., MASATO, G., GUILLOD, B. P. et al. (2016). Attributing human mortality during extreme heat waves to anthropogenic climate change. *Environ. Res. Lett.* **11** 074006.

OPITZ, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *J. Multivariate Anal.* **122** 409–413. MR3189331 https://doi.org/10.1016/j.jmva.2013.08.008

OPITZ, T. (2016). Modeling asymptotically independent spatial extremes based on Laplace random fields. *Spat. Stat.* **16** 1–18. MR3493085 https://doi.org/10.1016/j.spasta.2016.01.001

OSBORN, T. J. and BRIFFA, K. R. (2006). The spatial extent of 20th-century warmth in the context of the past 1200 years. *Science* **311** 841–844.

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

PADOAN, S. A. (2013). Extreme dependence models based on event magnitude. *J. Multivariate Anal.* **122** 1–19. MR3189305 https://doi.org/10.1016/j.jmva.2013.07.009

PADOAN, S. A., RIBATET, M. and SISSON, S. A. (2010). Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.* **105** 263–277. MR2757202 https://doi.org/10.1198/jasa.2009.tm08577

QUENOUILLE, M. H. (1949). Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B* **11** 68–84. MR0032176

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

REICH, B. J. and SHABY, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6** 1430–1451. MR3058670 https://doi.org/10.1214/12-AOAS591

RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer, New York. MR0900810 https://doi.org/10.1007/978-0-387-75953-1

SCARROTT, C. and MACDONALD, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT* **10** 33–60. MR2912370

SCHLATHER, M. (2002). Models for stationary max-stable random fields. *Extremes* **5** 33–44. MR1947786 https://doi.org/10.1023/A:1020977924878

SHOOTER, R., ROSS, E., TAWN, J. and JONATHAN, P. (2019). On spatial conditional extremes for ocean storm severity. *Environmetrics* **30** e2562, 18 pp. MR4009977 https://doi.org/10.1002/env.2562

THEURILLAT, J.-P. and GUISAN, A. (2001). Potential impact of climate change on vegetation in the European Alps: A review. *Clim. Change* **50** 77–109.

TUKEY, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Stat.* **29** 614.

VARIN, C., REID, N. and FIRTH, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42. MR2796852

WADSWORTH, J. L. and TAWN, J. A. (2012). Dependence modelling for spatial extremes. *Biometrika* **99** 253–272. MR2931252 https://doi.org/10.1093/biomet/asr080

WADSWORTH, J. L. and TAWN, J. A. (2019). Higher-dimensional spatial extremes via single-site conditioning. Preprint. Available at arXiv:1912.06560.

ZHONG, P., HUSER, R. and OPITZ, T. (2022). Supplement to "Modeling nonstationary temperature maxima based on extremal dependence changing with event magnitude." https://doi.org/10.1214/21-AOAS1504SUPPA, https://doi.org/10.1214/21-AOAS1504SUPPB

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 300–325
https://doi.org/10.1214/21-AOAS1505

SEQUENTIAL MODELING, MONITORING, AND FORECASTING OF STREAMING WEB TRAFFIC DATA

BY KAORU IRIE[1,a], CHRIS GLYNN[2,b] AND TEVFIK AKTEKIN[3,c]

[1]*Faculty of Economics, University of Tokyo,* [a]*irie@e.u-tokyo.ac.jp*
[2]*Economic Research, Zillow Group,* [b]*christophergl@zillowgroup.com*
[3]*Paul College of Business and Economics, University of New Hampshire,* [c]*tevfik.aktekin@unh.edu*

In this paper we introduce strategies for modeling, monitoring, and forecasting sequential web traffic data using flows from the Fox News website. In our analysis we consider a family of Poisson-gamma state space (PGSS) models that can accurately quantify the uncertainty exhibited by web traffic data, can provide fast sequential monitoring and prediction mechanisms for high frequency time intervals, and are computationally feasible when structural breaks are present. As such, we extend the family of PGSS models to include the state augmented (sa-)PGSS model whose state evolution structure is flexible and responsive to sudden changes. Such adaptability is achieved by augmenting the state vector of the PGSS model with an additional state variable for a time-varying discount factor. We develop an efficient particle-based estimation procedure that is suitable for sequential analysis, allowing us to estimate dynamic state variables and static parameters via closed-form conditional sufficient statistics. We compare the performance of the PGSS family of models against viable alternatives from the literature and argue that, especially in the presence of structural breaks, our proposed approach yields superior sequential model fit and predictive performance while preserving computational feasibility. We provide additional insights by designing a simulation study that mimics potential web traffic data patterns.

REFERENCES

AKTEKIN, T., POLSON, N. and SOYER, R. (2018). Sequential Bayesian analysis of multivariate count data. *Bayesian Anal.* **13** 385–409. MR3780428 https://doi.org/10.1214/17-BA1054

AKTEKIN, T. and SOYER, R. (2011). Call center arrival modeling: A Bayesian state-space approach. *Naval Res. Logist.* **58** 28–42. MR2796402 https://doi.org/10.1002/nav.20436

AKTEKIN, T., SOYER, R. and XU, F. (2013). Assessment of mortgage default risk via Bayesian state space models. *Ann. Appl. Stat.* **7** 1450–1473. MR3127954 https://doi.org/10.1214/13-AOAS632

BERRY, L. R., HELMAN, P. and WEST, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction–sales time series. *Int. J. Forecast.* **36** 552–569.

BERRY, L. R. and WEST, M. (2020). Bayesian forecasting of many count-valued time series. *J. Bus. Econom. Statist.* **38** 872–887. MR4154894 https://doi.org/10.1080/07350015.2019.1604372

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096 https://doi.org/10.1093/biomet/81.3.541

CARVALHO, C. M., JOHANNES, M. S., LOPES, H. F. and POLSON, N. G. (2010a). Particle learning and smoothing. *Statist. Sci.* **25** 88–106. MR2741816 https://doi.org/10.1214/10-STS325

CARVALHO, C. M., LOPES, H. F., POLSON, N. G. and TADDY, M. A. (2010b). Particle learning for general mixtures. *Bayesian Anal.* **5** 709–740. MR2740154 https://doi.org/10.1214/10-BA525

CHEN, X., BANKS, D. and WEST, M. (2019). Bayesian dynamic modeling and monitoring of network flows. *Netw. Sci.* **7** 292–318.

CHEN, X., IRIE, K., BANKS, D., HASLINGER, R., THOMAS, J. and WEST, M. (2018). Scalable Bayesian modeling, monitoring, and analysis of dynamic network flow data. *J. Amer. Statist. Assoc.* **113** 519–533. MR3832205 https://doi.org/10.1080/01621459.2017.1345742

DAVIS, R., HOLAN, S., LUND, R. and RAVISHANKER, N. (2015). *Handbook of Discrete-Valued Time Series*. CRC Press/CRC, Boca Raton.

Key words and phrases. Web traffic, count data, high frequency, Poisson-gamma, sequential Monte Carlo.

DOORNIK, J. A. (2007). *Object-Oriented Matrix Programming Using Ox*, 3rd ed. Timberlake Consultants Press and Oxford, London.

FEARNHEAD, P. (2002). Markov chain Monte Carlo, sufficient statistics, and particle filters. *J. Comput. Graph. Statist.* **11** 848–862. MR1951601 https://doi.org/10.1198/106186002321018821

FREELAND, R. K. and MCCABE, B. P. M. (2004). Analysis of low count time series data by Poisson autoregression. *J. Time Series Anal.* **25** 701–722. MR2089191 https://doi.org/10.1111/j.1467-9892.2004.01885.x

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889 https://doi.org/10.1111/j.1467-9892.1994.tb00184.x

FRÜHWIRTH-SCHNATTER, S. and WAGNER, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* **93** 827–841. MR2285074 https://doi.org/10.1093/biomet/93.4.827

GAMERMAN, D., REZENDE DOS SANTOS, T. and FRANCO, G. C. (2013). A non-Gaussian family of state-space models with exact marginal likelihood. *J. Time Series Anal.* **34** 625–645. MR3127211 https://doi.org/10.1111/jtsa.12039

GLYNN, C., TOKDAR, S. T., HOWARD, B. and BANKS, D. L. (2019). Bayesian analysis of dynamic linear topic models. *Bayesian Anal.* **14** 53–80. MR3910038 https://doi.org/10.1214/18-BA1100

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F* (*Radar and Signal Processing*) **140** 107–113. IET.

GRAMACY, R. B. and POLSON, N. G. (2011). Particle learning of Gaussian process models for sequential design and optimization. *J. Comput. Graph. Statist.* **20** 102–118. MR2816540 https://doi.org/10.1198/jcgs.2010.09171

HARVEY, A. C. and FERNANDES, C. (1989). Time series models for count or qualitative observations. *J. Bus. Econom. Statist.* **7** 407–417.

LOPES, H. F. and POLSON, N. G. (2016). Particle learning for fat-tailed distributions. *Econometric Rev.* **35** 1666–1691. MR3511035 https://doi.org/10.1080/07474938.2015.1092809

LOPES, H. F. and TSAY, R. S. (2011). Particle filters and Bayesian inference in financial econometrics. *J. Forecast.* **30** 168–209. MR2758809 https://doi.org/10.1002/for.1195

PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328 https://doi.org/10.2307/2670179

PRADO, R. and LOPES, H. F. (2013). Sequential parameter learning and filtering in structured autoregressive state-space models. *Stat. Comput.* **23** 43–57. MR3018349 https://doi.org/10.1007/s11222-011-9289-1

PRADO, R. and WEST, M. (2010). *Time Series*: *Modeling*, *Computation and Inference*. CRC Press, Boca Raton.

SINGPURWALLA, N. D., POLSON, N. G. and SOYER, R. (2018). From least squares to signal processing and particle filtering. *Technometrics* **60** 146–160. MR3804244 https://doi.org/10.1080/00401706.2017.1341341

SMITH, R. L. and MILLER, J. E. (1986). A non-Gaussian state space model and application to prediction of records. *J. Roy. Statist. Soc. Ser. B* **48** 79–88. MR0848053

STORVIK, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.* **50** 281–289.

UHLIG, H. (1994). On singular Wishart and singular multivariate beta distributions. *Ann. Statist.* **22** 395–405. MR1272090 https://doi.org/10.1214/aos/1176325375

UHLIG, H. (1997). Bayesian vector autoregressions with stochastic volatility. *Econometrica* **65** 59–73. MR1433685 https://doi.org/10.2307/2171813

WEST, M. and HARRISON, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Assoc.* **81** 741–750.

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1482232

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 326–348
https://doi.org/10.1214/21-AOAS1506

THE ROLE OF INTRINSIC DIMENSION IN HIGH-RESOLUTION PLAYER TRACKING DATA—INSIGHTS IN BASKETBALL

BY EDGAR SANTOS-FERNANDEZ[1,a], FRANCESCO DENTI[2,c], KERRIE MENGERSEN[1,b] AND ANTONIETTA MIRA[3,d]

[1]*School of Mathematical Sciences, Queensland University of Technology,* [a]*santosfe@qut.edu.au,* [b]*k.mengersen@qut.edu.au*
[2]*Department of Statistics, University of California, Irvine,* [c]*fdenti@uci.edu*
[3]*Faculty of Economics, Università della Svizzera italiana,* [d]*antonietta.mira@usi.ch*

Following the introduction of high-resolution player tracking technology, a new range of statistical analysis has emerged in sports, specifically in basketball. However, such high-dimensional data are often challenging for statistical inference and decision making. In this article we employ a state-of-the-art Bayesian mixture model that allows the estimation of heterogeneous intrinsic dimension (ID) within a dataset, and we propose some theoretical enhancements. Informally, the ID can be seen as an indicator of complexity and dependence of the data at hand, and it is usually assumed unique. Our method provides the capacity to reveal valuable insights about the hidden dynamics of sports interactions in space and time which helps to translate complex patterns into more coherent statistics. The application of this technique is illustrated using NBA basketball players' tracking data, allowing effective classification and clustering. In movement data the analysis identified key stages of offensive actions, such as creating space for passing, preparation/shooting, and following through which are relevant for invasion sports. We found that the ID value spikes, reaching a peak between four and eight seconds in the offensive part of the court, after which it declines. In shot charts we obtained groups of shots that produce substantially higher and lower successes. Overall, game-winners tend to have a larger intrinsic dimension, indicative of greater unpredictability and unique shot placements. Similarly, we found higher ID values in plays when the score margin is smaller rather than larger. The exploitation of these results can bring clear strategic advantages in sports games.

REFERENCES

ALLEGRA, M., FACCO, E., DENTI, F., LAIO, A. and MIRA, A. (2020). Data segmentation based on the local intrinsic dimension. *Sci. Rep.* **10** 16449.

ANSUINI, A., LAIO, A., MACKE, J. H. and ZOCCOLAN, D. (2019). Intrinsic dimension of data representations in deep neural networks.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969 https://doi.org/10.1214/aos/1176342871

BARTER, R. and YU, B. (2017). superheat: A graphical tool for exploring complex datasets using heatmaps. R package version 0.1.0.

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

BENNETT, R. S. (1969). The intrinsic dimensionality of signal collections. *IEEE Trans. Inf. Theory* **15** 517–525. https://doi.org/10.1109/TIT.1969.1054365

CALIŃSKI, T. and HARABASZ, J. (1974). A dendrite method for cluster analysis. *Commun. Stat.* **3** 1–27. MR0375641 https://doi.org/10.1080/03610927408827101

CAMASTRA, F. and STAIANO, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Inform. Sci.* **328** 26–41.

Key words and phrases. Bayesian clustering, high-dimensional data, intrinsic dimension, plays classification, movement data.

CAMPADELLI, P., CASIRAGHI, E., CERUTI, C. and ROZZA, A. (2015). Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Math. Probl. Eng.* Art. ID 759567, 21 pp. MR3417646 https://doi.org/10.1155/2015/759567

CELEUX, G. (1998). Bayesian inference for mixture: The label switching problem. *Compstat* 227–232. https://doi.org/10.1007/978-3-662-01131-7_26

CERVONE, D., D'AMOUR, A., BORNN, L. and GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *J. Amer. Statist. Assoc.* **111** 585–599. MR3538688 https://doi.org/10.1080/01621459.2016.1141685

D'AMOUR, A., CERVONE, D., BORNN, L. and GOLDSBERRY, K. (2015). Move or die: How ball movement creates open shots in the NBA. In *MIT Sloan Sports Analytics Conference*.

DE SILVA, V., CAINE, M., SKINNER, J., DOGAN, S., KONDOZ, A., PETER, T., AXTELL, E., BIRNIE, M. and SMITH, B. (2018). Player tracking data analytics as a tool for physical performance management in football: A case study from Chelsea football club academy. *Sports* **6** 130.

DODGE, S., BOHRER, G., WEINZIERL, R., DAVIDSON, S. C., KAYS, R., DOUGLAS, D., CRUZ, S., HAN, J., BRANDES, D. et al. (2013). The environmental-data automated track annotation (Env-DATA) system: Linking animal tracks with environmental data. *Mov. Ecol.* **1** 3.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510 https://doi.org/10.1080/01621459.1995.10476550

FACCO, E., D'ERRICO, M., RODRIGUEZ, A. and LAIO, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **7** 12140. https://doi.org/10.1038/s41598-017-11873-y

FRANKS, A., MILLER, A., BORNN, L. and GOLDSBERRY, K. (2015). Characterizing the spatial structure of defensive skill in professional basketball. *Ann. Appl. Stat.* **9** 94–121. MR3341109 https://doi.org/10.1214/14-AOAS799

FRITSCH, A. (2012). mcclust: Process an MCMC sample of clusterings. R package version 1.0.

FRÜHWIRTH-SCHNATTER, S. (2011). Dealing with label switching under model uncertainty. In *Mixtures: Estimation and Applications. Wiley Ser. Probab. Stat.* 213–239. Wiley, Chichester. MR2883354 https://doi.org/10.1002/9781119995678.ch10

GOLDSBERRY, K. (2012). Courtvision: New visual and spatial analytics for the NBA. In 2012 *MIT Sloan Sports Analytics Conference*.

HOBBS, W., MORGAN, S., GORMAN, A. D., MOONEY, M. and FREESTON, J. (2018). Playing unpredictably: Measuring the entropy of ball trajectories in international women's basketball. *Int. J. Perform. Anal. Sport* **18** 115–126.

LAMAS, L., BARRERA, J., OTRANTO, G. and UGRINOWITSCH, C. (2014). Invasion team sports: Strategy and match modeling. *Int. J. Perform. Anal. Sport* **14** 307–329.

LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. MR2351079 https://doi.org/10.1198/106186007X238855

LAZAR, N. (2014). The big picture: Take me out to the ball game. *Chance* **27** 45–48.

LEVINA, E. and BICKEL, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems* 777–784.

LUCEY, P., BIALKOWSKI, A., CARR, P., FOOTE, E. and MATTHEWS, I. A. (2012). Characterizing multi-agent team behavior from partial team tracings: Evidence from the English Premier League. In *AAAI*.

LUTZ, D. (2012). A cluster analysis of NBA players. In *MITSloan Sports Analytics Conference*.

MASTRANTONIO, G., GRAZIAN, C., MANCINELLI, S. and BIBBONA, E. (2019). New formulation of the logistic-Gaussian process to analyze trajectory tracking data. *Ann. Appl. Stat.* **13** 2483–2508. MR4037438 https://doi.org/10.1214/19-aoas1289

METULINI, R. (2018). Players movements and team shooting performance: A data mining approach for basketball. Preprint. Available at arXiv:1805.02501.

METULINI, R., MANISERA, M. and ZUCCOLOTTO, P. (2017). Space-time analysis of movements in basketball using sensor data. Preprint. Available at arXiv:1707.00883.

NISTALA, A. and GUTTAG, J. (2019). Using deep learning to understand patterns of player movement in the NBA. In *Proceedings of the MIT Sloan Sports Analytics Conference* 1–14.

PAPPALARDO, L., CINTIA, P., ROSSI, A., MASSUCCO, E., FERRAGINA, P., PEDRESCHI, D. and GIANNOTTI, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Sci. Data* **6** 1–15.

PEDERSEN, T. L. and ROBINSON, D. (2019). gganimate: A grammar of animated graphics. R package version 1.0.4.

PETRALIA, F., RAO, V. and DUNSON, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems* 1889–1897.

ROBERT, C. P. (2010). Multimodality and label switching: A discussion. In *Workshop on Mixtures, ICMS*.

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

ROSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65.

ROZZA, A., LOMBARDI, G., CERUTI, C., CASIRAGHI, E. and CAMPADELLI, P. (2012). Novel high intrinsic dimensionality estimators. *Mach. Learn.* **89** 37–65. MR2967955 https://doi.org/10.1007/s10994-012-5294-7

SAMPAIO, J., DRINKWATER, E. J. and LEITE, N. M. (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. *Eur. J. Sport Sci.* **10** 141–149.

SAMPAIO, J., MCGARRY, T., CALLEJA-GONZÁLEZ, J., SÁIZ, S. J., I DEL ALCÁZAR, X. S. and BALCIUNAS, M. (2015). Exploring game performance in the National Basketball Association using player tracking data. *PLoS ONE* **10** e0132894.

SANTOS-FERNANDEZ, E., DENTI, F., MENGERSEN, K. and MIRA, A. (2022). Supplement to "The role of intrinsic dimension in high-resolution player tracking data—Insights in basketball." https://doi.org/10.1214/21-AOAS1506SUPPA, https://doi.org/10.1214/21-AOAS1506SUPPB

SHORTRIDGE, A., GOLDSBERRY, K. and ADAMS, M. (2014). Creating space to shoot: Quantifying spatial relative field goal efficiency in basketball. *J. Quant. Anal. Sports* **10** 303–313. https://doi.org/10.1515/jqas-2013-0094

SKINNER, B. and GOLDMAN, M. (2017). Optimal strategy in basketball. In *Handbook of Statistical Methods and Analyses in Sports. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 229–244. CRC Press, Boca Raton, FL. MR3837238

SLOWIKOWSKI, K. (2019). ggrepel: Automatically position non-overlapping text labels with 'ggplot2'. R package version 0.8.1.

SPERRIN, M., JAKI, T. and WIT, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Stat. Comput.* **20** 357–366. MR2725393 https://doi.org/10.1007/s11222-009-9129-8

TERAMOTO, M., CROSS, C. L., RIEGER, R. H., MAAK, T. G. and WILLICK, S. E. (2018). Predictive validity of National Basketball Association draft combine on future performance. *J. Strength Cond. Res.* **32** 396–408.

VAN HAAREN, J., BEN SHITRIT, H., DAVIS, J. and FUA, P. (2016). Analyzing volleyball match data from the 2014 World Championships using machine learning techniques. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 627–634.

WADE, S. and GHAHRAMANI, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.* **13** 559–626. MR3807860 https://doi.org/10.1214/17-BA1073

WICKHAM, H. (2017). tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 349–367
https://doi.org/10.1214/21-AOAS1514

IN-GAME WIN PROBABILITIES FOR THE NATIONAL RUGBY LEAGUE

BY TIANYU GUAN[1,a], ROBERT NGUYEN[2,b], JIGUO CAO[3,c] AND TIM SWARTZ[3,d]

[1]*Department of Mathematics and Statistics, Brock University,* [a]*tguan@brocku.ca*
[2]*Department of Statistics, School of Mathematics and Statistics, University of New South Wales,* [b]*robert.nguyen@unsw.edu.au*
[3]*Department of Statistics and Actuarial Science, Simon Fraser University,* [c]*jiguo_cao@sfu.ca,* [d]*tim@stat.sfu.ca*

This paper develops new methods for providing instantaneous in-game win probabilities for the National Rugby League. Besides the score differential, betting odds, and real-time features extracted from the match event data are also used as inputs to inform the in-game win probabilities. Rugby matches evolve continuously in time, and the circumstances change over the duration of the match. Therefore, the match data are considered as functional data, and the in-game win probability is a function of the time of the match. We express the in-game win probability using a conditional probability formulation, the components of which are evaluated from the perspective of functional data analysis. Specifically, we model the score differential process and functional feature extracted from the match event data as sums of mean functions and noises. The mean functions are approximated by B-spline basis expansions with functional parameters. Since each match is conditional on a unique kickoff win probability of the home team obtained from the betting odds (i.e., the functional data are not independent and identically distributed), we propose a weighted least squares method to estimate the functional parameters by borrowing the information from matches with similar kickoff win probabilities. The variance and covariance elements are obtained by the maximum likelihood estimation method. The proposed method is applicable to other sports when suitable match event data are available.

REFERENCES

AINSWORTH, L. M., ROUTLEDGE, R. and CAO, J. (2011). Functional data analysis in ecosystem research: The decline of Oweekeno Lake sockeye salmon and Wannock River flow. *J. Agric. Biol. Environ. Stat.* **16** 282–300. MR2818550 https://doi.org/10.1007/s13253-010-0049-z

ALBERT, J., GLICKMAN, M. E., SWARTZ, T. B. and KONING, R. H., eds. (2017). *Handbook of Statistical Methods and Analyses in Sports. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3838291 https://doi.org/10.1201/9781315166070

BESSE, P. and RAMSAY, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika* **51** 285–311. MR0848110 https://doi.org/10.1007/BF02293986

BOOTH, M. and ORR, R. (2017). Time-loss injuries in sub-elite and emerging rugby league players. *J. Sports Sci. Med.* **16** 295–301.

BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138 https://doi.org/10.1007/978-1-4612-1154-9

BUTTREY, S. E., WASHBURN, A. R. and PRICE, W. L. (2011). Estimating NHL scoring rates. *J. Quant. Anal. Sports* **7** 1–18.

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496 https://doi.org/10.1214/009053606000000830

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

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

CERVONE, D., D'AMOUR, A., BORNN, L. and GOLDSBERRY, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. *J. Amer. Statist. Assoc.* **111** 585–599. MR3538688 https://doi.org/10.1080/01621459.2016.1141685

Key words and phrases. In-game win probability, event data, NRL, functional data analysis, model validation.

CHEN, T. and FAN, Q. (2018). A functional data approach to model score difference process in professional basketball games. *J. Appl. Stat.* **45** 112–127. MR3736861 https://doi.org/10.1080/02664763.2016.1268106

CLAUSET, A., KOGAN, M. and REDNER, S. (2015). Safe leads and lead changes in competitive team sports. *Phys. Rev. E* (3) **91** 062815, 11. MR3491426 https://doi.org/10.1103/PhysRevE.91.062815

DE BOOR, C. (2001). *A Practical Guide to Splines*, Revised ed. *Applied Mathematical Sciences* **27**. Springer, New York. MR1900298

DELAIGLE, A. and HALL, P. (2012). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 267–286. MR2899863 https://doi.org/10.1111/j.1467-9868.2011.01003.x

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. Springer, New York. MR2229687

GABBETT, T. J. (2005). Science of rugby league football: A review. *J. Sports Sci.* **23** 961–976.

GABLE, A. and REDNER, S. (2012). Random walk picture of basketball scoring. *J. Quant. Anal. Sports* **8** 1–20.

GLASSBROOK, D. J., DOYLE, T. L. A., ALDERSON, J. A. and FULLER, J. T. (2019). The demands of professional rugby league match-play: A meta-analysis. *Sports Medicine—Open* **5** Article number: 24.

GUAN, T., NGUYEN, R., CAO, J. and SWARTZ, T. (2022). Supplement to "In-game win probabilities for the National Rugby League." https://doi.org/10.1214/21-AOAS1514SUPP

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 https://doi.org/10.1214/009053606000000957

HASTIE, T. and MALLOWS, C. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 140–143.

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

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

JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408. MR1995716 https://doi.org/10.1198/016214503000189

KAYHAN, V. O. and WATKINS, A. (2018). A data snapshot approach for making real-time predictions in basketball. *Big Data* **6** 96–112. https://doi.org/10.1089/big.2017.0054

KAYHAN, V. O. and WATKINS, A. (2019). Predicting the point spread in professional basketball in real time: A data snapshot approach. *Journal of Business Analytics* **2** 63–73.

KING, T., JENKINS, D. and GABBETT, T. (2009). A time-motion analysis of professional rugby league match-play. *J. Sports Sci.* **27** 213–219.

KOKOSZKA, P. and REIMHERR, M. (2017). *Introduction to Functional Data Analysis*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3793167

LEE, A. (1999). Applications: Modelling rugby league data via bivariate negative binomial regression. *Aust. N. Z. J. Stat.* **14** 141–152.

LENG, X. and MÜLLER, H.-G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22** 68–76. https://doi.org/10.1093/bioinformatics/bti742

LOCK, D. and NETTLETON, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *J. Quant. Anal. Sports* **10** 197–205.

LUO, W., CAO, J., GALLAGHER, M. and WILES, J. (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Stat. Med.* **32** 2681–2694. MR3067415 https://doi.org/10.1002/sim.5684

MORRIS, J. S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.

PARMAR, N., JAMES, N., HUGHES, M., JONES, H. and HEARNE, G. (2017). Team performance indicators that predict match outcome and points difference in professional rugby league. *International Journal of Performance Analysis in Sport* **17** 1044–1056.

PETTIGREW, S. (2015). Assessing the offensive productivity of NHL players using in-game win probabilities. In *Proceedings of the* 9*th MIT Sloan Sports Analytics Conference*.

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

ROBBERECHTS, P., VAN HAAREN, J. and DAVIS, J. (2019). Who will win it? An in-game win probability model for football. In *Proceedings of the* 6*th Workshop on Machine Learning and Data Mining for Sports Analytics*, 20 *September* 2019 page 13, Würzburg, Germany.

SEITZ, L. B., RIVIÈRE, M., DE VILLARREAL, E. S. and HAFF, G. G. (2014). The athletic performance of elite rugby league players is improved after an 8-week small-sided game training intervention. *J. Strength Cond. Res.* **28** 971–975.

SONG, K., GAO, Y. and SHI, J. (2020). Making real-time predictions for NBA basketball games by combining the historical data and bookmaker's betting line. *Phys. A* **547** 124411.

STERN, H. S. (1994). A Brownian motion model for the progress of sports scores. *J. Amer. Statist. Assoc.* **89** 1128–1134.

ŠTRUMBELJ, E. and VRAČAR, P. (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. *Int. J. Forecast.* **28** 532–542.

VRAČAR, P., ŠTRUMBELJ, E. and KONONENKO, I. (2016). Modeling basketball play-by-play data. *Expert Syst. Appl.* **44** 58–66.

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

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Review of functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WINDT, J., GABBETT, T. J., FERRIS, D. and KHAN, K. M. (2017). Training load–injury paradox: Is greater preseason participation associated with lower in-season injury risk in elite rugby league players? *Br. J. Sports Med.* **51** 645–650. https://doi.org/10.1136/bjsports-2016-095973

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561 https://doi.org/10.1198/016214504000001745

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

YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857 https://doi.org/10.1214/09-AOS772

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 368–390
https://doi.org/10.1214/21-AOAS1480

MANIFOLD VALUED DATA ANALYSIS OF SAMPLES OF NETWORKS, WITH APPLICATIONS IN CORPUS LINGUISTICS

BY KATIE E. SEVERN[a], IAN L. DRYDEN[c] AND SIMON P. PRESTON[b]

School of Mathematical Sciences, University of Nottingham, [a]*katie.severn@nottingham.ac.uk,*
[b]*simon.preston@nottingham.ac.uk*

Department of Mathematics and Statistics, Florida International University, [c]*ian.dryden@fiu.edu*

Networks arise in many applications, such as in the analysis of text documents, social interactions and brain activity. We develop a general framework for extrinsic statistical analysis of samples of networks, motivated by networks representing text documents in corpus linguistics. We identify networks with their graph Laplacian matrices for which we define metrics, embeddings, tangent spaces and a projection from Euclidean space to the space of graph Laplacians. This framework provides a way of computing means, performing principal component analysis, regression, and carrying out hypothesis tests, such as for testing for equality of means between two samples of networks. We apply the methodology to the set of novels by Jane Austen and Charles Dickens.

REFERENCES

AMARAL, G. J. A., DRYDEN, I. L. and WOOD, A. T. A. (2007). Pivotal bootstrap methods for k-sample problems in directional statistics and shape analysis. *J. Amer. Statist. Assoc.* **102** 695–707. MR2370861 https://doi.org/10.1198/016214506000001400

ANDERSON, E. (2018). rosqp: Quadratic programming solver using the OSQP library. R package version 0.1.0.

BAKKER, C., HALAPPANAVAR, M. and SATHANUR, A. V. (2018). Dynamic graphs, community detection, and Riemannian geometry. *Appl. Netw. Sci.* **3** 3.

BHATTACHARYA, R. and LIN, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Amer. Math. Soc.* **145** 413–428. MR3565392 https://doi.org/10.1090/proc/13216

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498 https://doi.org/10.1214/aos/1046294456

BHATTACHARYA, R. and PATRANGENARU, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.* **33** 1225–1259. MR2195634 https://doi.org/10.1214/009053605000000093

CHARLES DICKENS INFO (2020). Charles Dickens timeline. Available at https://www.charlesdickensinfo.com/life/timeline/.

COOTES, T. F., TAYLOR, C. J., COOPER, D. H. and GRAHAM, J. (1994). Image search using flexible shape models generated from sets of examples. In *Statistics and Images*: *Vol.* 2 (K. V. Mardia, ed.) 111–139. Carfax, Oxford.

DE KLERK, E. (2002). *Aspects of Semidefinite Programming*: *Interior Point Algorithms and Selected Applications*. *Applied Optimization* **65**. Kluwer Academic, Dordrecht. MR2064921 https://doi.org/10.1007/b105286

DRYDEN, I. L. (2019). `shapes` package. R Foundation for Statistical Computing, Vienna, Austria. Contributed package, Version 1.2.5.

DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. MR2750388 https://doi.org/10.1214/09-AOAS249

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3559734 https://doi.org/10.1002/9781119072492

DRYDEN, I. L., PENNEC, X. and PEYRAT, J.-M. (2010). Power Euclidean metrics for covariance matrices with application to diffusion tensor imaging. arXiv e-prints. Available at arXiv:1009.3045.

EVERT, S. (2008). Corpora and collocations. *Corpus Linguistics. An International Handbook* **2** 1212–1248.

FLETCHER, P. T., LU, C., PIZER, S. M. and JOSHI, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.* **23** 995–1005.

Key words and phrases. Extrinsic mean, graph Laplacian, regression, Riemannian, hypothesis test.

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464

FU, A., NARASIMHAN, B., KANG, D. W., DIAMOND, S. and MILLER, J. (2020). CVXR: disciplined convex optimization. R package version 1.0-1.

GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. MR3693544 https://doi.org/10.1214/16-AOAS1015

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 427–444. MR2155347 https://doi.org/10.1111/j.1467-9868.2005.00510.x

HIGHAM, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR0943997 https://doi.org/10.1016/0024-3795(88)90223-6

KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1891212 https://doi.org/10.1002/9780470317006

KENT, J. T. (1994). The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc. Ser. B* **56** 285–299. MR1281934

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

KOLACZYK, E. D., LIN, L., ROSENBERG, S., WALTERS, J. and XU, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *Ann. Statist.* **48** 514–538. MR4065172 https://doi.org/10.1214/19-AOS1820

LE, H. L. (1995). Mean size-and-shapes and mean shapes: A geometric point of view. *Adv. in Appl. Probab.* **27** 44–55. MR1315576 https://doi.org/10.2307/1428094

LIN, L., ST. THOMAS, B., ZHU, H. and DUNSON, D. B. (2017). Extrinsic local regression on manifold-valued data. *J. Amer. Statist. Assoc.* **112** 1261–1273. MR3735375 https://doi.org/10.1080/01621459.2016.1208615

MAHLBERG, M., STOCKWELL, P., DE JOODE, J., SMITH, C. and O'DONNELL, M. B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* **11** 433–463.

MASAROTTO, V., PANARETOS, V. M. and ZEMEL, Y. (2019). Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. *Sankhya A* **81** 172–213. MR3982195 https://doi.org/10.1007/s13171-018-0130-1

PHILLIPS, M. K. (1983). Lexical macrostructure in science text. Ph.D. thesis.

PIGOLI, D., ASTON, J. A. D., DRYDEN, I. L. and SECCHI, P. (2014). Distances and inference for covariance operators. *Biometrika* **101** 409–422. MR3215356 https://doi.org/10.1093/biomet/asu008

PRESTON, S. P. and WOOD, A. T. A. (2010). Two-sample bootstrap hypothesis tests for three-dimensional labelled landmark data. *Scand. J. Stat.* **37** 568–587. MR2779637 https://doi.org/10.1111/j.1467-9469.2010.00690.x

R CORE TEAM (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

ROCKAFELLAR, R. T. (1993). Lagrange multipliers and optimality. *SIAM Rev.* **35** 183–238. MR1220880 https://doi.org/10.1137/1035044

SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32. MR2183942 https://doi.org/10.2202/1544-6115.1175

SEVERN, K. E, DRYDEN, I. L and PRESTON, S. P (2022). Supplement to "Manifold valued data analysis of samples of networks, with applications in corpus linguistics." https://doi.org/10.1214/21-AOAS1480SUPP

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. and IDEKER, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13** 2498–2504.

SHAW, N. (1990). Free indirect speech and Jane Austen's 1816 revision of Northanger Abbey. *Studies in English Literature*, 1500–1900 **30** 591–601.

THE JANE AUSTEN SOCIETY OF NORTH AMERICA (2020). Jane Austen's works. Available at http://jasna.org/austen/works/.

VILLANI, C. (2009). *Optimal Transport*: *Old and New*. Springer, Berlin.

WARD, J. H. JR. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58** 236–244. MR0148188

WILKS, S. S. (1962). *Mathematical Statistics*. *A Wiley Publication in Mathematical Statistics*. Wiley, New York. MR0144404

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 391–413
https://doi.org/10.1214/21-AOAS1496

MULTIVARIATE MIXED MEMBERSHIP MODELING: INFERRING DOMAIN-SPECIFIC RISK PROFILES

BY MASSIMILIANO RUSSO[1,a], BURTON H. SINGER[2,b] AND DAVID B. DUNSON[3,c]

[1]*Harvard-MIT Center for Regulatory Science, Harvard Medical School,* [a]*m_russo@hms.harvard.edu*
[2]*Emerging Pathogens Institute and Department of Mathematics, University of Florida,* [b]*bhsinger@epi.ufl.edu*
[3]*Department of Statistical Science, Duke University,* [c]*dunson@duke.edu*

Characterizing the shared memberships of individuals in a classification scheme poses severe interpretability issues, even when using a moderate number of classes (say four). Mixed membership models quantify this phenomenon, but they typically focus on goodness-of-fit more than on interpretable inference. To achieve a good numerical fit, these models may, in fact, require many extreme profiles, making the results difficult to interpret. We introduce a new class of multivariate mixed membership models that, when variables can be partitioned into subject-matter based domains, can provide a good fit to the data using fewer profiles than standard formulations. The proposed model explicitly accounts for the blocks of variables corresponding to the distinct domains along with a cross-domain correlation structure which provides new information about shared membership of individuals in a complex classification scheme. We specify a multivariate logistic normal distribution for the membership vectors which allows easy introduction of auxiliary information leveraging a latent multivariate logistic regression. A Bayesian approach to inference, relying on Pólya gamma data augmentation, facilitates efficient posterior computation via Markov chain Monte Carlo. We apply this methodology to a spatially explicit study of malaria risk over time on the Brazilian Amazon frontier.

REFERENCES

AIROLDI, E., BLEI, D., XING, E. and FIENBERG, S. (2005). A latent mixed membership model for relational data. In *Proceedings of the* 3*rd International Workshop on Link Discovery* (*LinkKDD'*05). ACM, New York, NY, USA, 82–89.

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

AIROLDI, E. M., BLEI, D. M., EROSHEVA, E. A. and FIENBERG, S. E., eds. (2015) *Handbook of Mixed Membership Models and Their Applications. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3381023

AITCHISON, J. and SHEN, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67** 261–272. MR0581723 https://doi.org/10.2307/2335470

ARTIN, M. (1991). *Algebra*. Prentice Hall, Englewood Cliffs, NJ. MR1129886

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

BANERJEE, S., GELFAND, A. E. and POLASEK, W. (2000). Geostatistical modelling for spatial interaction data with application to postal service performance. *J. Statist. Plann. Inference* **90** 87–105. MR1791583 https://doi.org/10.1016/S0378-3758(00)00111-7

BERKMAN, L., SINGER, B. and MANTON, K. (1989). Black/white differences in health status and mortality among the elderly. *Demography* **26** 661–678.

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. MR2949366 https://doi.org/10.1080/01621459.2011.646934

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

Key words and phrases. Admixture model, contingency table, latent Dirichlet allocation, multivariate categorical data, multivariate logistic normal distribution.

CASTRO, M. C., SAWYER, D. O. and SINGER, B. H. (2007). Spatial patterns of malaria in the Amazon: Implications for surveillance and targeted interventions. *Health Place* **13** 368–380.
CASTRO, M. C., MONTE-MÓR, R. L., SAWYER, D. O. and SINGER, B. H. (2006). Malaria risk on the Amazon frontier. *Proc. Natl. Acad. Sci. USA* **103** 2452–2457. https://doi.org/10.1073/pnas.0510576103
CHEN, Y.-C., WANG, Y. S. and EROSHEVA, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. *Ann. Appl. Stat.* **12** 846–876. MR3834288 https://doi.org/10.1214/18-AOAS1169
CHUIT, R., GURTLER, R. E., MAC DOUGALL, L., SEGURA, E. L. and SINGER, B. (2001). Chagas disease-risk assessment by an environmental approach in northern Argentina. *Rev. Patol. Trop.* **30** 193–208.
EROSHEVA, E. A. and FIENBERG, S. E. (2005). Bayesian mixed membership models for soft clustering and classification. In *Classification — the Ubiquitous Challenge* (C. Weihs and W. Gaul, eds.) 11–26. Springer, Berlin, Heidelberg.
EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745 https://doi.org/10.1214/07-AOAS126
EROSHEVA, E., FIENBERG, S. and LAFFERTY, J. (2004). Mixed-membership models of scientific publications. *Proc. Natl. Acad. Sci. USA* **101** 5220–5227. https://doi.org/10.1073/pnas.0307760101
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
GETIS, A. and ORD, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **24** 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x
GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235. https://doi.org/10.1073/pnas.0307752101
GROSS, J. H. and MANRIQUE-VALLIER, D. (2015). A mixed membership approach to the assessment of political ideology from survey responses. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 119–139. CRC Press, Boca Raton, FL. MR3380027
KAO, E. K., SMITH, S. T. and AIROLDI, E. M. (2019). Hybrid mixed-membership blockmodel for inference on realistic network interactions. *IEEE Trans. Netw. Sci. Eng.* **6** 336–350. MR4014178 https://doi.org/10.1109/tnse.2018.2823324
KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X
LAFFERTY, J. D. and BLEI, D. M. (2006). Correlated topic models. *Adv. Neural Inf. Process. Syst.* **18** 147–154.
LINDERMAN, S., JOHNSON, M. and ADAMS, R. P. (2015). Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Adv. Neural Inf. Process. Syst.* **28** 3456–3464.
MILLER, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **63** 81–97.
PAGANIN, S., HERRING, A. H., OLSHAN, A. F., DUNSON, D. B. and STUDY, T. N. B. D. P. (2021). Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Anal.* **16** 301–370. Includes comments and discussions by 19 discussants and a rejoinder by the authors. MR4255332 https://doi.org/10.1214/20-ba1197
POLSON, N. G. and SCOTT, J. G. (2011). Default Bayesian analysis for multi-way tables: A data-augmentation approach. Preprint. Available at arXiv:1109.4180.
POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712 https://doi.org/10.1080/01621459.2013.829001
RUSSO, M., SINGER, B. H. and DUNSON, D. B. (2022). Supplement to "Multivariate mixed membership modeling: Inferring domain-specific risk profiles." https://doi.org/10.1214/21-AOAS1496SUPPA, https://doi.org/10.1214/21-AOAS1496SUPPB
SINGER, B. (1989). Grade of membership representations: Concepts and problems. In *Probability, Statistics, and Mathematics* 317–334. Academic Press, Boston, MA. MR1031295
SINGER, B. H. and CASTRO, M. C. (2015). Interpretability constraints and trade-offs in using mixed membership models. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 159–172. CRC Press, Boca Raton, FL. MR3380029
SMITH, A. N. and ALLENBY, G. M. (2020). Demand models with random partitions. *J. Amer. Statist. Assoc.* **115** 47–65. MR4078444 https://doi.org/10.1080/01621459.2019.1604360
STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293 https://doi.org/10.1111/1467-9868.00265
WADE, R. H. (2011). Boulevard of broken dreams: The inside story of the World Bank's Polonoroeste Road Project in Brazil's Amazon GRI Working Papers No. 55, Grantham Research Institute on Climate Change and the Environment.

WANG, Y. S. and EROSHEVA, E. A. (2015). mixedMem: Tools for Discrete Multivariate Mixed Membership Models. R package version 1.1.0.

WOODBURY, M. A., CLIVE, J. and GARSON, A. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Comput. Biomed. Res.* **11** 277–298.

WORLD BANK (1992). World Bank Approaches to the Environment in Brazil, Vol. V: The Polonoroeste Program. OECD Report 10039, Sec M92-64.

XU, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Statist.* **45** 675–707. MR3650397 https://doi.org/10.1214/16-AOS1464

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 414–435
https://doi.org/10.1214/21-AOAS1498

ESTIMATING THE EFFECTIVENESS OF PERMANENT PRICE REDUCTIONS FOR COMPETING PRODUCTS USING MULTIVARIATE BAYESIAN STRUCTURAL TIME SERIES MODELS

BY FIAMMETTA MENCHETTI[1,a] AND IAVOR BOJINOV[2,b]

[1]*DiSIA, Universitá di Firenze,* [a]*fiammetta.menchetti@unifi.it*
[2]*Technology and Operations Management, Harvard Business School,* [b]*ibojinov@hbs.edu*

The Florence branch of an Italian supermarket chain recently implemented a strategy that permanently lowered the price of numerous store brands in several product categories. To quantify the impact of such a policy change, researchers often use synthetic control methods for estimating causal effects when a subset of units receive a single persistent treatment and the rest are unaffected by the change. In our applications, however, competitor brands not assigned to treatment are likely impacted by the intervention because of substitution effects; more broadly, this type of interference occurs whenever the treatment assignment of one unit affects the outcome of another. This paper extends the synthetic control methods to accommodate partial interference, allowing interference within predefined groups but not between them. Focusing on a class of causal estimands that capture the effect both on the treated and control units, we develop a multivariate Bayesian structural time series model for generating synthetic controls that would have occurred in the absence of an intervention, enabling us to estimate our novel effects. In a simulation study we explore our Bayesian procedures' empirical properties and show that it achieves good frequentists coverage, even when the model is misspecified. We use our new methodology to make causal statements about the impact on sales of the affected store brands and their direct competitors. Our proposed approach is implemented in the CausalMBSTS R package.

REFERENCES

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929 https://doi.org/10.1198/jasa.2009.ap08746

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2015). Comparative politics and the synthetic control method. *Amer. J. Polit. Sci.* **59** 495–510.

ABADIE, A. and GARDEAZABAL, J. (2003). The economic costs of conflict: A case study of the Basque country. *Am. Econ. Rev.* **93** 113–132.

BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2018). The augmented synthetic control method. Preprint. Available at arXiv:1811.04170.

BILLMEIER, A. and NANNICINI, T. (2013). Assessing economic liberalization episodes: A synthetic control approach. *Rev. Econ. Stat.* **95** 983–1001.

BLATTBERG, R. C., BRIESCH, R. and FOX, E. J. (1995). How promotions work. *Mark. Sci.* **14** G122–G132.

BOJINOV, I., CHEN, A. and LIU, M. (2020). The importance of being causal. *Harvard Data Science Review*.

BOJINOV, I. and MENCHETTI, F. (2020). CausalMBSTS: MBSTS Models for Causal Inference and Forecasting. R package version 0.1.0.

BOJINOV, I., RAMBACHAN, A. and SHEPHARD, N. (2020). Panel Experiments and Dynamic Causal Effects: A Finite Population Perspective. Preprint. Available at arXiv:2003.09915.

BOJINOV, I. and SHEPHARD, N. (2019). Time series experiments and causal estimands: Exact randomization tests and trading. *J. Amer. Statist. Assoc.* **114** 1665–1682. MR4047291 https://doi.org/10.1080/01621459.2018.1527225

Key words and phrases. Causal inference, partial interference, synthetic controls, Bayesian structural time Series.

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. MR3341115 https://doi.org/10.1214/14-AOAS788

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 627–641. MR1626005 https://doi.org/10.1111/1467-9868.00144

CAO, J. and DOWD, C. (2019). Estimation and inference for synthetic control methods with spillover effects. Preprint. Available at arXiv:1902.07343.

COX, D. R. (1958). *Planning of Experiments. A Wiley Publication in Applied Statistics*. Wiley, New York. MR0095561

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265

DUBE, A. and ZIPPERER, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies. IZA Discussion Paper 8944.

FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. *J. Amer. Statist. Assoc.* **116** 901–918. MR4270033 https://doi.org/10.1080/01621459.2020.1768100

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*. CRC press, Boca Raton.

GOBILLON, L. and MAGNAC, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Rev. Econ. Stat.* **98** 535–551.

GROSSI, G., LATTARULO, P., MARIANI, M., MATTEI, A. and ÖNER, Ö. (2020). Synthetic Control Group Methods in the Presence of Interference: The Direct and Spillover Effects of Light Rail on Neighborhood Retail Activity. Preprint. Available at arXiv:2004.05027.

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.

KEOGH, E. and RATANAMAHATANA, C. A. (2005). Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **7** 358–386.

KREIF, N., GRIEVE, R., HANGARTNER, D., TURNER, A. J., NIKOLOVA, S. and SUTTON, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Econ.* **25** 1514–1528. https://doi.org/10.1002/hec.3258

LARSEN, K. (2019). MarketMatching Package Vignette. R package version 1.1.2.

LI, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *J. Amer. Statist. Assoc.* **115** 2068–2083. MR4189777 https://doi.org/10.1080/01621459.2019.1686986

MENCHETTI, F. and BOJINOV, I. (2022). Supplement to "Estimating the effectiveness of permanent price reductions for competing products using multivariate Bayesian structural time series models." https://doi.org/10.1214/21-AOAS1498SUPPA, https://doi.org/10.1214/21-AOAS1498SUPPB

NESLIN, S. A., HENDERSON, C. and QUELCH, J. (1985). Consumer promotions and the acceleration of product purchases. *Mark. Sci.* **4** 147–165.

NICHOLSON, W. and SNYDER, C. M. (2012). *Microeconomic Theory: Basic Principles and Extensions*. Nelson Education.

O'NEILL, S., KREIF, N., GRIEVE, R., SUTTON, M. and SEKHON, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Serv. Outcomes Res. Methodol.* **16** 1–21. https://doi.org/10.1007/s10742-016-0146-8

PAPADOGEORGOU, G., MEALLI, F., ZIGLER, C. M., DOMINICI, F., WASFY, J. H. and CHOIRAT, C. (2018). Causal Impact of the Hospital Readmissions Reduction Program on Hospital Readmissions and Mortality. Preprint. Available at arXiv:1809.09590.

PAUWELS, K., HANSSENS, D. M. and SIDDARTH, S. (2002). The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *J. Mark. Res.* **39** 421–439.

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., GREENLAND, S. and HU, F.-C. (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *J. Amer. Statist. Assoc.* **94** 687–712. MR1723276 https://doi.org/10.2307/2669978

ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112

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

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

SALVADOR, S. and CHAN, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* **11** 561–580.

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

VANDERWEELE, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociol. Methods Res.* **38** 515–544. MR2758165 https://doi.org/10.1177/0049124110366236

VIVIANO, D. and BRADIC, J. (2019). Synthetic learner: Model-free inference on treatments over time. Preprint. Available at arXiv:1904.01490.

WEST, M. and HARRISON, J. (2006). *Bayesian Forecasting and Dynamic Models*. Springer, Berlin.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trab. Estad. Investig. Oper.* **31** 585–603.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 436–459
https://doi.org/10.1214/21-AOAS1499

BAYESIAN ADJUSTMENT FOR PREFERENTIAL TESTING IN ESTIMATING INFECTION FATALITY RATES, AS MOTIVATED BY THE COVID-19 PANDEMIC

BY HARLAN CAMPBELL[1,a], PERRY DE VALPINE[2], LAUREN MAXWELL[3], VALENTIJN M. T. DE JONG[4], THOMAS P. A. DEBRAY[4,5], THOMAS JAENISCH[3,6] AND PAUL GUSTAFSON[1,b]

[1]*Department of Statistics, University of British Columbia,* [a]*harlan.campbell@stat.ubc.ca,* [b]*gustaf@stat.ubc.ca*
[2]*Department of Environmental Science, Policy, and Management, University of California*
[3]*Heidelberg Institute for Global Health, Heidelberg University Hospital*
[4]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University*
[5]*Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University*
[6]*Department of Epidemiology, Colorado School of Public Health*

A key challenge in estimating the infection fatality rate (IFR), along with its relation with various factors of interest, is determining the total number of cases. The total number of cases is not known not only because not everyone is tested but also, more importantly, because tested individuals are not representative of the population at large. We refer to the phenomenon whereby infected individuals are more likely to be tested than noninfected individuals as "preferential testing." An open question is whether or not it is possible to reliably estimate the IFR without any specific knowledge about the degree to which the data are biased by preferential testing. In this paper we take a partial identifiability approach, formulating clearly where deliberate prior assumptions can be made and presenting a Bayesian model which pools information from different samples. When the model is fit to European data obtained from seroprevalence studies and national official COVID-19 statistics, we estimate the overall COVID-19 IFR for Europe to be 0.53%, 95% C.I. = [0.38%, 0.70%].

REFERENCES

ANDERSON, R. M., HEESTERBEEK, H., KLINKENBERG, D. and HOLLINGSWORTH, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **395** 931–934.

BENDAVID, E., MULANEY, B., SOOD, N., SHAH, S., LING, E., BROMLEY-DULFANO, R., LAI, C., WEISSBERG, Z., SAAVEDRA, R. et al. (2020). COVID-19 antibody seroprevalence in Santa Clara County, California. medRxiv.

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

BIRRELL, P. J., DE ANGELIS, D. and PRESANIS, A. M. (2018). Evidence synthesis for stochastic epidemic models. *Statist. Sci.* **33** 34–43. MR3757502 https://doi.org/10.1214/17-STS631

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662 https://doi.org/10.2307/1390675

CAMPBELL, H., DE JONG, V. M., MAXWELL, L., DEBRAY, T., JAENISCH, T. and GUSTAFSON, P. (2020). Measurement error in meta-analysis (MEMA)–a Bayesian framework for continuous outcome data. *Res. Synth. Methods*. https://doi.org/10.1002/jrsm.1515.

CAMPBELL, H., DE VALPINE, P., MAXWELL, L., DE JONG, V. M., DEBRAY, T. P., JAENISCH, T. and GUSTAFWSON, P. (2022). Supplement to "Bayesian adjustment for preferential testing in estimating infection fatality rates, as motivated by the COVID-19 pandemic." https://doi.org/10.1214/21-AOAS1499SUPP

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**.

Key words and phrases. Selection bias, partial identification, evidence synthesis.

COCHRAN, J. J. (2020). Why we need more coronavirus tests than we think. *Significance* 14–15.
DE ANGELIS, D., PRESANIS, A. M., BIRRELL, P. J., TOMBA, G. S. and HOUSE, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10** 83–87.
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
FAUST, J. S. (2020). Comparing COVID-19 deaths to flu deaths is like comparing apples to oranges. *Sci. Am.*.
FOG, A. (2008). Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Comm. Statist. Simulation Comput.* **37** 241–257. MR2422884 https://doi.org/10.1080/03610910701790236
GELMAN, A. and CARPENTER, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 1269–1283. MR4166866
GELMAN, A., RUBIN, D. B. et al. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
GREENLAND, S. (2005). Multiple-bias modelling for analysis of observational data. *J. Roy. Statist. Soc. Ser. A* **168** 267–306. MR2119402 https://doi.org/10.1111/j.1467-985X.2004.00349.x
GREWELLE, R. and DE LEO, G. (2020). Estimating the global infection fatality rate of COVID-19. medRxiv.
GUDBJARTSSON, D. F., HELGASON, A., JONSSON, H., MAGNUSSON, O. T., MELSTED, P., NORDDAHL, G. L., SAEMUNDSDOTTIR, J., SIGURDSSON, A., SULEM, P. et al. (2020). Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.*
GUSTAFSON, P. (2010). Bayesian inference for partially identified models. *Int. J. Biostat.* **6** 17. MR2602560 https://doi.org/10.2202/1557-4679.1206
GUSTAFSON, P. and GREENLAND, S. (2009). Interval estimation for messy observational data. *Statist. Sci.* **24** 328–342. MR2757434 https://doi.org/10.1214/09-STS305
HALE, T., WEBSTER, S., PETHERICK, A., PHILLIPS, T. and KIRA, B. (2020). Oxford COVID-19 government response tracker. Available at https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker#data.
HAUSER, A., COUNOTTE, M. J., MARGOSSIAN, C. C., KONSTANTINOUDIS, G., LOW, N., ALTHAUS, C. L. and RIOU, J. (2020). Estimation of SARS-CoV-2 mortality during the early stages of an epidemic: A modelling study in Hubei, China and northern Italy. *PLoS Med.* **17** e1003189.
IOANNIDIS, J. (2020a). The infection fatality rate of COVID-19 inferred from seroprevalence data. (version 2 (June 8, 2020–14:00)). medRxiv.
IOANNIDIS, J. P. (2020b). First Opinion: A fiasco in the making? as the coronavirus pandemic takes hold, we are making decisions without reliable data. STAT. 2020. Available at https://tinyurl.com/uj539o4.
KOBAYASHI, T., JUNG, S.-M., LINTON, N. M., KINOSHITA, R., HAYASHI, K., MIYAMA, T., ANZAI, A., YANG, Y., YUAN, B. et al. (2020). Communicating the risk of death from novel coronavirus disease (COVID-19). *J. Clin. Med.* **9**. https://doi.org/10.3390/jcm9020580
KRUSCHKE, J. (2014). *Doing Bayesian Data Analysis*: *A Tutorial with R*, *JAGS*, *and Stan*. Academic Press, San Diego.
KÜMMERER, M., BERENS, P. and MACKE, J. (2020). A simple Bayesian analysis of the infection fatality rate in Gangelt, and an uncertainty aware extrapolation to infection-counts in Germany. Available at https://matthias-k.github.io/BayesianHeinsberg.html.
LAMBERT, P. C., SUTTON, A. J., BURTON, P. R., ABRAMS, K. R. and JONES, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat. Med.* **24** 2401–2428. MR2151713 https://doi.org/10.1002/sim.2112
LEE, G. (2020). Coronavirus: Why so many people are dying in Belgium. BBC.com. Available at https://www.bbc.com/news/world-europe-52491210.
LEON, D. A., SHKOLNIKOV, V. M., SMEETH, L., MAGNUS, P., PECHHOLDOVÁ, M. and JARVIS, C. I. (2020). COVID-19: A need for real-time monitoring of weekly excess deaths. *Lancet* **395** e81.
LI, S. and HUA, X. (2020). The closer to the Europe Union headquarters, the higher risk of COVID-19? Cautions regarding ecological studies of COVID-19. medRxiv.
LINTON, N. M., KOBAYASHI, T., YANG, Y., HAYASHI, K., AKHMETZHANOV, A. R., JUNG, S.-M., YUAN, B., KINOSHITA, R. and NISHIURA, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *J. Clin. Med.* **9** 538.
LIPSITCH, M. (2020). First Opinion: We know enough now to act decisively against COVID-19. social distancing is a good place to start. STAT. Available at https://tinyurl.com/yx4gf9mr.
LYONS, N. (1980). Closed expressions for noncentral hypergeometric probabilities. *Comm. Statist. Simulation Comput.* **9** 313–314.
MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. *Springer Series in Statistics*. Springer, New York. MR2151380

MEYEROWITZ-KATZ, G. and MERONE, L. (2020). A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates (version 4). medRxiv.

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

NEIL, M., FENTON, N., OSMAN, M. and MCLACHLAN, S. (2020). Bayesian network analysis of COVID-19 data reveals higher infection prevalence rates and lower fatality rates than widely reported. medRxiv.

O'DRISCOLL, M., DOS SANTOS, G. R., WANG, L., CUMMINGS, D. A., AZMAN, A. S., PAIREAU, J., FONTANET, A., CAUCHEMEZ, S. and SALJE, H. (2020). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* 1–6.

ONDER, G., REZZA, G. and BRUSAFERRO, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *JAMA* **323** 1775–1776. https://doi.org/10.1001/jama.2020.4683

OWID (2020). Codebook for the complete our world in data COVID-19 dataset. Available at https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data-codebook.md.

PASTOR-BARRIUSO, R., PÉREZ-GÓMEZ, B., HERNÁN, M. A., PÉREZ-OLMEDA, M., YOTTI, R., OTEO-IGLESIAS, J., SANMARTÍN, J. L., LEÓN-GÓMEZ, I., FERNÁNDEZ-GARCÍA, A. et al. (2020). Infection fatality risk for SARS-CoV-2 in community dwelling population of Spain: Nationwide seroepidemiological study. *BMJ* **371**.

PEARCE, N. (2000). The ecological fallacy strikes back. *J. Epidemiol. Community Health* **54** 326–327.

PELLIS, L., CAUCHEMEZ, S., FERGUSON, N. M. and FRASER, C. (2020). Systematic selection between age and household structure for models aimed at emerging epidemic predictions. *Nat. Commun.* **11** 1–11.

POLLÁN, M., PÉREZ-GÓMEZ, B., PASTOR-BARRIUSO, R., OTEO, J., HERNÁN, M. A., PÉREZ-OLMEDA, M., SANMARTÍN, J. L., FERNÁNDEZ-GARCÍA, A., CRUZ, I. et al. (2020). Prevalence of SARS-CoV-2 in Spain (ENE-COVID): A nationwide, population-based seroepidemiological study. *Lancet*.

PRESANIS, A. M., DE ANGELIS, D., THE NEW YORK CITY SWINE FLU INVESTIGATION TEAM, HAGY, A., REED, C., RILEY, S., COOPER, B. S., FINELLI, L. et al. (2009). The severity of pandemic H1N1 influenza in the United States. *PLoS Med.* **6**.

PROCHASKA, C. and THEODORE, L. (2018). Discrete probability distributions. *Introd. Math. Methods Environ. Eng. Sci.* 287.

RINALDI, G. and PARADISI, M. (2020). An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. medRxiv.

SAHAI, H. and KHURSHID, A. (1995). *Statistics in Epidemiology*: *Methods*, *Techniques and Applications*. CRC press, Boca Raton.

SMEDT, T. D., MERRALL, E., MACINA, D., PEREZ-VILAR, S., ANDREWS, N. and BOLLAERTS, K. (2018). Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. *PLoS ONE* **13** e0199180. https://doi.org/10.1371/journal.pone.0199180

STEVENS, W. L. (1951). Mean and variance of an entry in a contingency table. *Biometrika* **38** 468–470. MR0047287 https://doi.org/10.1093/biomet/38.3-4.468

STREECK, H., SCHULTE, B., KUEMMERER, B., RICHTER, E., HÖLLER, T., FUHRMANN, C., BARTOK, E., DOLSCHEID, R., BERGER, M. et al. (2020). Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *Nat. Commun.* **11** 1–12.

SUTTON, A. J., COOPER, N. J., JONES, D. R., LAMBERT, P. C., THOMPSON, J. R. and ABRAMS, K. R. (2007). Evidence-based sample size calculations based upon updated meta-analysis. *Stat. Med.* **26** 2479–2500. MR2364400 https://doi.org/10.1002/sim.2704

THOMPSON, S. G. and HIGGINS, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* **21** 1559–1573.

WONG, J. Y., HEATH KELLY, D. K., WU, J. T., LEUNG, G. M. and COWLING, B. J. (2013). Case fatality risk of influenza A (H1N1pdm09): A systematic review. *Epidemiology* **24**.

WU, J. T., LEUNG, K., BUSHMAN, M., KISHORE, N., NIEHUS, R., DE SALAZAR, P. M., COWLING, B. J., LIPSITCH, M. and LEUNG, G. M. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26** 506–510. https://doi.org/10.1038/s41591-020-0822-7

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 460–476
https://doi.org/10.1214/21-AOAS1507

PREELECTORAL POLLS VARIABILITY: A HIERARCHICAL BAYESIAN MODEL TO ASSESS THE ROLE OF HOUSE EFFECTS WITH APPLICATION TO ITALIAN ELECTIONS

BY DOMENICO DE STEFANO[1,a], FRANCESCO PAULI[2,b] AND NICOLA TORELLI[2,c]

[1]*Department of Political and Social Science, University of Trieste,* [a]*ddestefano@units.it*
[2]*Department of Business, Economics, Mathematics, and Statistics, University of Trieste,* [b]*francesco.pauli@deams.units.it,* [c]*nicola.torelli@deams.units.it*

It is widely known that preelectoral polls often suffer from nonsampling errors that pollsters try to compensate for in final estimates by means of diverse ad hoc adjustments, thus leading to well-known house effects. We propose a Bayesian hierarchical model to investigate the role of house effects on the total variability of predictions. To illustrate the model, data from preelectoral polls in Italy in 2006, 2008 and 2013 are considered. Unlike alternative techniques or models, our proposal leads: (i) to correctly decompose the different sources of variability; (ii) to recognize the role of house effects; (iii) to evaluate its dynamics, showing that variability of house effects across pollsters diminishes as the date of election approaches; (iv) to investigate the relationship between house effects and overall prediction errors.

REFERENCES

AAPOR (2017). An Evaluation of 2016 Election Polls in the US. Available at https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx.

BLUMENTHAL, M. (2008). More on the 'Convergence Mystery'. Available at http://www.pollster.com/blogs/more_on_the_convergence_myster.php?nr=1, accessed 2018-07-10.

BLUMENTHAL, M. (2014). Polls, forecasts, and aggregators. *PS Polit. Sci. Polit.* **47** 297–300.

CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

DE STEFANO, D., PAULI, F. and TORELLI, N. (2022). Supplement to "Preelectoral polls variability: A hierarchical Bayesian model to assess the role of house effects with application to Italian elections." https://doi.org/10.1214/21-AOAS1507SUPP

DURAND, C. (2008). The polls of the 2007 French presidential campaign: Were lessons learned from the 2002 catastrophe? *Int. J. Public Opin. Res.* **20** 275–298.

ERIKSON, R. S., PANAGOPOULOS, C. and WLEZIEN, C. (2004). Likely (and unlikely) voters and the assessment of campaign dynamics. *Public Opin. Q.* **68** 588–601.

ERIKSON, R. S. and WLEZIEN, C. (1999). Presidential polls as a time series: The case of 1996. *Public Opin. Q.* **63** 163–177.

GAETAN, C. and GRIGOLETTO, M. (2004). Smoothing sample extremes with dynamic models. *Extremes* **7** 221–236. MR2143941 https://doi.org/10.1007/s10687-005-6474-7

GASPERONI, G. and CALLEGARO, M. (2008). Un miglioramento immeritato? La capacità predittiva dei sondaggi preelettorali e le elezioni politiche del 2008. *Polis* **22** 483–506.

GELMAN, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Stat. Rev.* **71** 369–382.

GELMAN, A. (2005). Analysis of variance—why it is more important than ever. *Ann. Statist.* **33** 1–53. MR2157795 https://doi.org/10.1214/009053604000001048

GELMAN, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electron. J. Stat.* **7** 2595–2602. MR3121624 https://doi.org/10.1214/13-EJS854

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

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Taylor & Francis, London.

Key words and phrases. Finite population variance, posterior predictive check, ANOVA, preelectoral polls.

HANRETTY, C. (2013). The 2013 Italian Election: A Poll-Based Forecast SSRN Scholarly Paper No. ID 2214605 Social Science Research Network.

HILLYGUS, D. S. (2011). The evolution of election polling in the United States. *Public Opin. Q.* **75** 962–981.

JACKMAN, S. (2005). Pooling the polls over an election campaign. *Aust. J. Polit. Sci.* **40** 499–517.

LAVRAKAS, P., TRAUGOTT, M., BLUM, M., ZUKIN, C. and DRESSER, D. (2008). The experts reply on the poll convergence mystery. *Surv. Pract.* **1**.

LINZER, D. A. (2012). Pollsters May Be Herding. Available at http://votamatic.org/pollsters-may-be-herding/, accessed 2018-07-10.

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

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2006). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=09/04/2006&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2008). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=13/04/2008&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MINISTERO DELL'INTERNO, UFFICIO IV—SERVIZI INFORMATICI ELETTORALI (2013). Archivio storico delle elezioni. Available at http://elezionistorico.interno.it/index.php?tpel=C&dtel=24/02/2013&tpa=I&tpe=A&lev0=0&levsut0=0&es0=S&ms=S, accessed 2018-06-03.

MOORE, D. (2008). Evaluating the 2008 pre-election polls—the convergence mystery. *Surv. Pract.* **1**.

PANAGOPOULOS, C. (2009). Polls and elections: Preelection poll accuracy in the 2008 general elections. *Pres. Stud. Q.* **39** 896–907.

PASEK, J. (2015). Predicting elections: Considering tools to pool the polls. *Public Opin. Q.* **79** 594–619.

PICKUP, M. and JOHNSTON, R. (2007). Campaign trial heats as electoral information: Evidence from the 2004 and 2006 Canadian federal elections. *Elect. Stud.* **26** 460–476.

PICKUP, M. and JOHNSTON, R. (2008). Campaign trial heats as election forecasts: Measurement error and bias in 2004 presidential campaign polls. *Int. J. Forecast.* **24** 272–284.

PRESIDENZA DEL CONSIGLIO DEI MINISTRI—DIPARTIMENTO PER L'INFORMAZIONE E L'EDITORIA (2015). Il Sito Ufficiale dei Sondaggi Politici ed Elettorali. Available at http://www.sondaggipoliticoelettorali.it/, accessed 2018-07-10.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

SILVER, N. (2010). Pollster Ratings V4.0: Methodology. Available at http://www.fivethirtyeight.com/2010/06/pollster-ratings-v40-methodology.html, accessed 2018-07-10.

SPECKMAN, P. L. and SUN, D. (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika* **90** 289–302. MR1986647 https://doi.org/10.1093/biomet/90.2.289

STAN DEVELOPMENT TEAM (2016). RStan: The R interface to Stan.

STURGIS, P., BAKER, N., CALLEGARO, M., FISHER, S., GREEN, J., JENNINGS, W., KUHA, J., LAUDERDALE, B. and SMITH, P. (2016). Report of the inquiry into the 2015 British general election opinion polls.

WLEZIEN, C. and ERIKSON, R. S. (2007). The horse race: What polls reveal as the election campaign unfolds. *Int. J. Public Opin. Res.* **19** 74–88.

WORCESTER, R. (1996). Political polling: 95% expertise and 5% luck. *J. Roy. Statist. Soc. Ser. A* **159** 5–20.

YUE, Y. R., SPECKMAN, P. L. and SUN, D. (2012). Priors for Bayesian adaptive spline smoothing. *Ann. Inst. Statist. Math.* **64** 577–613. MR2880870 https://doi.org/10.1007/s10463-010-0321-6

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 477–494
https://doi.org/10.1214/21-AOAS1509

DETECTING AND MODELING CHANGES IN A TIME SERIES OF PROPORTIONS

BY THOMAS J. FISHER[1,a], JING ZHANG[1,b], STEPHEN P. COLEGATE[2,c] AND MICHAEL J. VANNI[3,d]

[1]*Department of Statistics, Miami University,* [a]*fishert4@miamioh.edu,* [b]*zhangj8@miamioh.edu*

[2]*Division of Biostatistics & Bioinformatics, Department of Environmental & Public Health Sciences, University of Cincinnati,* [c]*colegasn@mail.uc.edu*

[3]*Department of Biology, Miami University,* [d]*vannimj@miamioh.edu*

We propose a framework to detect and model shifts in a time series of continuous proportions, that is, a vector of proportions measuring the parts of a whole. By reparameterizing the shape of a Dirichlet distribution, we can model the location and scale separately through generalized linear models. A hidden Markov model allows the coefficients of the generalized linear models to change, thus allowing for the time series to undergo multiple regimes. This framework allows a practitioner to adequately model seasonality, trends, or include covariate information as well as detect change points. The model's behavior is studied via simulation and through the analysis of lake phytoplankton data from 1992 through 2012. Our analyses demonstrate that the model can be effective in detecting and modeling changes in a time series of proportions. Pertaining to the phytoplankton data, the overall biomass has grown with some changes to the community level dynamics occurring circa 2000. Specifically, the proportion of cyanobacteria appears to have increased to the detriment of diatoms.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. With discussion. MR0676206

AITCHISON, J. (1985). A general class of distributions on the simplex. *J. Roy. Statist. Soc. Ser. B* **47** 136–146. MR0805071

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

AUE, A. and HORVÁTH, L. (2013). Structural breaks in time series. *J. Time Series Anal.* **34** 1–16. MR3008012 https://doi.org/10.1111/j.1467-9892.2012.00819.x

BARCELÓ-VIDAL, C., AGUILAR, L. and MARTÍN-FERNÁNDEZ, J. A. (2011). Compositional VARIMA Time Series. In *Compositional Data Analysis* 87–103. Wiley, New York.

BARDWELL, L. and FEARNHEAD, P. (2017). Bayesian detection of abnormal segments in multiple time series. *Bayesian Anal.* **12** 193–218. MR3597572 https://doi.org/10.1214/16-BA998

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

BINDER, K. E., POURAHMADI, M. and MJELDE, J. W. (2018). The role of temporal dependence in factor selection and forecasting oil prices. *Empir. Econ.* 1–39.

BRUNSDON, T. M. and SMITH, T. M. F. (1998). The time series analysis of compositional data. *J. Off. Stat.* **14** 237–253.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer, New York. MR2159833

CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **41** 389–405.

CHIB, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86** 221–241. MR1649222 https://doi.org/10.1016/S0304-4076(97)00115-2

Key words and phrases. Change points, compositional data, Dirichlet regression, hidden Markov model, time series.

ERDMAN, C. and EMERSON, J. W. (2008). A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics* **24** 2143–2148.

FEARNHEAD, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput.* **16** 203–213. MR2227396 https://doi.org/10.1007/s11222-006-8450-8

FEARNHEAD, P. and LIU, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 589–605. MR2370070 https://doi.org/10.1111/j.1467-9868.2007.00601.x

FISHER, T. J., ZHANG, J., COLEGATE, S. P. and VANNI, M. J. (2022). Supplement to "Detecting and modeling changes in a time series of proportions." https://doi.org/10.1214/21-AOAS1509SUPPA, https://doi.org/10.1214/21-AOAS1509SUPPB

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

GRUNWALD, G. K., RAFTERY, A. E. and GUTTORP, P. (1993). Time series of continuous proportions. *J. Roy. Statist. Soc. Ser. B* **55** 103–116.

HAMPTON, S. E., HOLMES, E. E., SCHEEF, L. P., SCHEUERELL, M. D., KATZ, S. L., PENDLETON, D. E. and WARD, E. J. (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (MAR) models. *Ecology* **94** 2663–2669.

HAYES, N. M. and VANNI, M. J. (2018). Microcystin concentrations can be predicted with phytoplankton biomass and watershed morphology. *Inland Waters* **8** 273–283.

HIJAZI, R. H. and JERNIGAN, R. W. (2009). Modeling compositional data using Dirichlet regression models. *J. Appl. Probab. Stat.* **4** 77–91. MR2668780

HOLMES, M., KOJADINOVIC, I. and QUESSY, J.-F. (2013). Nonparametric tests for change-point detection à la Gombay and Horváth. *J. Multivariate Anal.* **115** 16–32. MR3004542 https://doi.org/10.1016/j.jmva.2012.10.004

JAMES, N. A. and MATTESON, D. S. (2014). Ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software* **62** 1–25.

KANG, S., LIU, G., QI, H. and WANG, M. (2018). Bayesian variance changepoint detection in linear models with symmetric heavy-tailed errors. *Comput. Econ.* **52** 459–477.

KELLY, P. T., GONZÁLEZ, M. J., RENWICK, W. H. and VANNI, M. J. (2018). Increased light availability and nutrient cycling by fish provide resilience against reversing eutrophication in an agriculturally impacted reservoir. *Limnol. Oceanogr.* **63** 2647–2660.

KOJADINOVIC, I. (2020). npcp: Some Nonparametric CUSUM Tests for Change-Point Detection in Possibly Multivariate Observations. R package version 0.2-0.

KOOP, G. and KOROBILIS, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Found Trends Econom.* **3** 267–358.

LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** 127–143. MR1145463 https://doi.org/10.1016/0304-4149(92)90141-C

LIANG, Z., QIAN, S. S., WU, S., CHEN, H., LIU, Y., YU, Y. and YI, X. (2019). Using Bayesian change point model to enhance understanding of the shifting nutrients-phytoplankton relationship. *Ecol. Model.* **393** 120–126.

LUND, R. and REEVES, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model. *J. Climate* **15** 2547–2554.

LUONG, T. M., PERDUCA, V. and NUEL, G. (2012). Hidden Markov model applications in change-point analysis. In *Hidden Markov Models–Applications in Signal, Image and Pattern Recognition*.

LYSTIG, T. C. and HUGHES, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *J. Comput. Graph. Statist.* **11** 678–689. MR1938450 https://doi.org/10.1198/106186002402

MATTESON, D. S. and JAMES, N. A. (2014b). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567 https://doi.org/10.1080/01621459.2013.849605

MATTESON, D. S. and TSAY, R. S. (2011). Dynamic orthogonal components for multivariate time series. *J. Amer. Statist. Assoc.* **106** 1450–1463. MR2896848 https://doi.org/10.1198/jasa.2011.tm10616

MILLS, T. C. (2010). Forecasting compositional time series. *Qual. Quant.* **44** 673–690.

O'REILLY, C. M., SHARMA, S., GRAY, D. K., HAMPTON, S. E., READ, J. S., ROWLEY, R. J., SCHNEIDER, P., LENTERS, J. D., MCINTYRE, P. B. et al. (2015). Rapid and highly variable warming of lake surface waters around the globe. *Geophys. Res. Lett.* **42** 10,773–10,781.

OHIO SUPERCOMPUTER CENTER (2016). Owens Supercomputer.

PAERL, H. W., OTTEN, T. G. and KUDELA, R. (2018). Mitigating the expansion of harmful algal blooms across the freshwater-to-marine continuum. *Environ. Sci. Technol.* **52** 5519–5529. PMID: 29656639.

PAWLOWSKY-GLAHN, V. and BUCCIANTI, A., eds. (2011). *Compositional Data Analysis: Theory and Applications* Wiley, Chichester. MR2920574 https://doi.org/10.1002/9781119976462

PEARSON, K. (1897). Mathematical contributions to the theory of evolution. – On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. R. Soc. Lond.* **60** 489–498.

PRABUCHANDRAN, K. J., SINGH, N., DAYAMA, P. and PANDIT, V. (2021). Change point detection for compositional multivariate data. *Appl. Intell.*.

RENWICK, W. H., VANNI, M. J., FISHER, T. J. and MORRIS, E. L. (2018). Stream nitrogen, phosphorus, and sediment concentrations show contrasting long-term trends associated with agricultural change. *J. Environ. Qual.* **47** 1513–1521. https://doi.org/10.2134/jeq2018.04.0162

ROBBINS, M. W., GALLAGHER, C. M. and LUND, R. B. (2016). A general regression changepoint test for time series data. *J. Amer. Statist. Assoc.* **111** 670–683. MR3538696 https://doi.org/10.1080/01621459.2015.1029130

ROBBINS, M., GALLAGHER, C., LUND, R. and AUE, A. (2011a). Mean shift testing in correlated data. *J. Time Series Anal.* **32** 498–511. MR2835683 https://doi.org/10.1111/j.1467-9892.2010.00707.x

ROBBINS, M. W., LUND, R. B., GALLAGHER, C. M. and LU, Q. (2011b). Changepoints in the North Atlantic tropical cyclone record. *J. Amer. Statist. Assoc.* **106** 89–99. MR2816704 https://doi.org/10.1198/jasa.2011.ap10023

ROBERT, C. P., CELEUX, G. and DIEBOLT, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statist. Probab. Lett.* **16** 77–83. MR1208503 https://doi.org/10.1016/0167-7152(93)90127-5

STAN DEVELOPMENT TEAM (2018). RStan: The R interface to Stan. R package version 2.18.2.

STEPHENS, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *J. Roy. Statist. Soc. Ser. C* **43** 159–178.

TSAY, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2778591 https://doi.org/10.1002/9780470644560

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

VEHTARI, A., GABRY, J., YAO, Y. and GELMAN, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.0.0.

VISSER, I. and SPEEKENBRINK, M. (2010). DepmixS4: An R package for hidden Markov models. *J. Stat. Softw.* **36** 1–21.

WEST, M. (2020). Bayesian forecasting of multivariate time series: Scalability, structure uncertainty and decisions. *Ann. Inst. Statist. Math.* **72** 1–31. MR4052647 https://doi.org/10.1007/s10463-019-00741-3

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 495–520
https://doi.org/10.1214/21-AOAS1510

PREDICTION OF HEREDITARY CANCERS USING NEURAL NETWORKS

BY ZOE GUAN[1,a], GIOVANNI PARMIGIANI[2,b], DANIELLE BRAUN[3,d] AND LORENZO TRIPPA[2,c]

[1]*Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center,* [a]*guanz@mskcc.org*
[2]*Department of Data Sciences, Dana-Farber Cancer Institute,* [b]*gp@jimmy.harvard.edu,* [c]*ltrippa@ds.dfci.harvard.edu*
[3]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [d]*dbraun@mail.harvard.edu*

Family history is a major risk factor for many types of cancer. Mendelian risk prediction models translate family histories into cancer risk predictions, based on knowledge of cancer susceptibility genes. These models are widely used in clinical practice to help identify high-risk individuals. Mendelian models leverage the entire family history, but they rely on many assumptions about cancer susceptibility genes that are either unrealistic or challenging to validate, due to low mutation prevalence. Training more flexible models, such as neural networks, on large databases of pedigrees can potentially lead to accuracy gains. In this paper we develop a framework to apply neural networks to family history data and investigate their ability to learn inherited susceptibility to cancer. While there is an extensive literature on neural networks and their state-of-the-art performance in many tasks, there is little work applying them to family history data. We propose adaptations of fully-connected neural networks and convolutional neural networks to pedigrees. In data simulated under Mendelian inheritance, we demonstrate that our proposed neural network models are able to achieve nearly optimal prediction performance. Moreover, when the observed family history includes misreported cancer diagnoses, neural networks are able to outperform the Mendelian BRCAPRO model embedding the correct inheritance laws. Using a large dataset of over 200,000 family histories, the Risk Service cohort, we train prediction models for future risk of breast cancer. We validate the models using data from the Cancer Genetics Network.

REFERENCES

AMUNDADOTTIR, L. T., THORVALDSSON, S., GUDBJARTSSON, D. F., SULEM, P., KRISTJANSSON, K., ARNASON, S., GULCHER, J. R., BJORNSSON, J., KONG, A. et al. (2004). Cancer as a complex phenotype: Pattern of cancer distribution within and beyond the nuclear family. *PLoS Med.* **1** e65.

ANTON-CULVER, H., ZIOGAS, A., BOWEN, D., FINKELSTEIN, D., GRIFFIN, C., HANSON, J., ISAACS, C., KASTEN-SPORTES, C., MINEAU, G. et al. (2003). The cancer genetics network: Recruitment results and pilot studies. *Publ. Health Genomics* **6** 171–177.

ANTONIOU, A. C., PHAROAH, P. D. P., MCMULLAN, G., DAY, N. E., STRATTON, M. R., PETO, J., PONDER, B. J. and EASTON, D. F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **86** 76–83. https://doi.org/10.1038/sj.bjc.6600008

ANTONIOU, A. C., PHAROAH, P. P. D., SMITH, P. and EASTON, D. F. (2004). The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br. J. Cancer* **91** 1580.

BALMAÑA, J., STOCKWELL, D. H., STEYERBERG, E. W., STOFFEL, E. M., DEFFENBAUGH, A. M., REID, J. E., WARD, B., SCHOLL, T., HENDRICKSON, B. et al. (2006). Prediction of MLH1 and MSH2 mutations in Lynch syndrome. *JAMA* **296** 1469–1478.

BANEGAS, M. P., JOHN, E. M., SLATTERY, M. L., GOMEZ, S. L., YU, M., LACROIX, A. Z., PEE, D., CHLEBOWSKI, R. T., HINES, L. M. et al. (2017). Projecting individualized absolute invasive breast cancer risk in US Hispanic women. *J. Natl. Cancer Inst.* **109** djw215.

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

Key words and phrases. Machine learning, Mendelian risk prediction, family history.

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

BERRY, D. A., PARMIGIANI, G., SANCHEZ, J., SCHILDKRAUT, J. and WINER, E. (1997). Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J. Natl. Cancer Inst.* **89** 227–237.

BERRY, D. A., IVERSEN JR., E. S., GUDBJARTSSON, D. F., HILLER, E. H., GARBER, J. E., PESHKIN, B. N., LERMAN, C., WATSON, P., LYNCH, H. T. et al. (2002). BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J. Clin. Oncol.* **20** 2701–2712.

BISHOP, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univ. Press, New York. MR1385195

BISWAS, S., ATIENZA, P., CHIPMAN, J., HUGHES, K., BARRERA, A. M. G., AMOS, C. I., ARUN, B. and PARMIGIANI, G. (2013). Simplifying clinical use of the genetic risk prediction model BRCAPRO. *Breast Cancer Res. Treat.* **139** 571–579.

BRAUN, D., GORFINE, M., KATKI, H. A., ZIOGAS, A., ANTON-CULVER, H. and PARMIGIANI, G. (2014). Extending Mendelian risk prediction models to handle misreported family history.

BRAUN, D., GORFINE, M., KATKI, H. A., ZIOGAS, A. and PARMIGIANI, G. (2018). Nonparametric adjustment for measurement error in time-to-event data: Application to risk prediction models. *J. Amer. Statist. Assoc.* **113** 14–25. MR3803436 https://doi.org/10.1080/01621459.2017.1311261

BRENTNALL, A. R., COHN, W. F., KNAUS, W. A., YAFFE, M. J., CUZICK, J. and HARVEY, J. A. (2019). A case-control study to add volumetric or clinical mammographic density into the Tyrer–Cuzick breast cancer risk model. *J. Breast Imaging* **1** 99–106.

CANNON-ALBRIGHT, L. A., CARR, S. R. and AKERLEY, W. (2019). Population-based relative risks for lung cancer based on complete family history of lung cancer. *J. Thorac. Oncol.* **14** 1184–1191. https://doi.org/10.1016/j.jtho.2019.04.019

CARAYOL, J. and BONAÏTI-PELLIÉ, C. (2004). Estimating penetrance from family data using a retrospective likelihood when ascertainment depends on genotype and age of onset. *Genet. Epidemiol.* **27** 109–117.

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

CASTALDI, P. J., DAHABREH, I. J. and IOANNIDIS, J. P. A. (2011). An empirical assessment of validation practices for molecular classifiers. *Brief. Bioinform.* **12** 189–202. https://doi.org/10.1093/bib/bbq073

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

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

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

CHEN, J., BAE, E., ZHANG, L., HUGHES, K., PARMIGIANI, G., BRAUN, D. and REBBECK, T. R. (2020). Penetrance of breast and ovarian cancer in women who carry a BRCA1/2 mutation and do not use risk-reducing salpingo-oophorectomy: An updated meta-analysis. *JNCI Cancer Spectr.* **4** pkaa029. https://doi.org/10.1093/jncics/pkaa029

CHIPMAN, J., DROHAN, B., BLACKFORD, A., PARMIGIANI, G., HUGHES, K. and BOSINOFF, P. (2013). Providing access to risk prediction tools via the HL7 XML-formatted risk web service. *Breast Cancer Res. Treat.* **140** 187–193.

CHOI, J., DEKKERS, O. M. and LE CESSIE, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur. J. Epidemiol.* **34** 23–36. https://doi.org/10.1007/s10654-018-0447-z

CHOI, Y.-H., KOPCIUK, K. A. and BRIOLLAIS, L. (2008). Estimating disease risk associated with mutated genes in family-based designs. *Hum. Hered.* **66** 238–251.

CHOUDHURY, P. P., MAAS, P., WILCOX, A., WHEELER, W., BROOK, M., CHECK, D., GARCIA-CLOSAS, M. and CHATTERJEE, N. (2020a). iCARE: An R package to build, validate and apply absolute risk models. *PLoS ONE* **15** e0228198.

CHOUDHURY, P. P., WILCOX, A. N., BROOK, M. N., ZHANG, Y., AHEARN, T., ORR, N., COULSON, P., SCHOEMAKER, M. J., JONES, M. E. et al. (2020b). Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *J. Natl. Cancer Inst.* **112** 278–285.

CLAUS, E. B., RISCH, N. and THOMPSON, W. D. (1994). Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. *Cancer* **73** 643–651.

CLEVERT, D.-A., UNTERTHINER, T. and HOCHREITER, S. (2015). Fast and accurate deep network learning by exponential linear units (ELUs). Preprint. Available at arXiv:1511.07289.

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

DONG, C. and HEMMINKI, K. (2001). Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. *Int. J. Cancer* **92** 144–150.

DONG, Y., SU, H., ZHU, J. and BAO, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. Preprint. Available at arXiv:1708.05493.

EASTON, D. F. (1999). How many more breast cancer predisposition genes are there? *Breast Cancer Res.* **1** 14.

EASTON, D. F., FORD, D. and BISHOP, D. T. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **56** 265–271.

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

EUHUS, D. M., SMITH, K. C., ROBINSON, L., STUCKY, A., OLOPADE, O. I., CUMMINGS, S., GARBER, J. E., CHITTENDEN, A., MILLS, G. B. et al. (2002). Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *J. Natl. Cancer Inst.* **94** 844–851.

FAN, F., XIONG, J. and WANG, G. (2020). On interpretability of artificial neural networks. Available at https://arxiv.org/abs/2001.02522.

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.

GAIL, M. H., COSTANTINO, J. P., PEE, D., BONDY, M., NEWMAN, L., SELVAN, M., ANDERSON, G. L., MALONE, K. E., MARCHBANKS, P. A. et al. (2007). Projecting individualized absolute invasive breast cancer risk in African American women. *J. Natl. Cancer Inst.* **99** 1782–1792.

GARCÍA-LAENCINA, P. J., SANCHO-GÓMEZ, J. and FIGUEIRAS-VIDAL, A. R. (2010). Pattern classification with missing data: A review. *Neural Comput. Appl.* **19** 263–282.

GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* **4** 1–58.

GERDS, T. A. and SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** 1029–1040. MR2312613 https://doi.org/10.1002/bimj.200610301

GINSBURG, G. S., WU, R. R. and ORLANDO, L. A. (2019). Family health history: Underused for actionable risk assessment. *Lancet* **394** 596–603.

GUAN, Z., PARMIGIANI, G., BRAUN, D. and TRIPPA, L. (2022a). Supplement to "Prediction of hereditary cancers using neural networks." https://doi.org/10.1214/21-AOAS1510SUPPA

GUAN, Z., PARMIGIANI, G., BRAUN, D. and TRIPPA, L. (2022b). Supplement to "Prediction of hereditary cancers using neural networks." https://doi.org/10.1214/21-AOAS1510SUPPB

HAMPSHIRE II, J. B. and PEARLMUTTER, B. (1991). Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In *Connectionist Models* 159–172. Elsevier, Amsterdam.

HECHTLINGER, Y., CHAKRAVARTI, P. and QIN, J. (2017). A generalization of convolutional neural networks to graph-structured data. Preprint. Available at arXiv:1704.08165.

HINTON, G., DENG, L., YU, D., DAHL, G. E., MOHAMED, A., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **29** 82–97.

HORNER, M. J., RIES, L. A. G., KRAPCHO, M., NEYMAN, N., AMINOU, R., HOWLADER, N., ALTEKRUSE, S. F., FEUER, E. J., HUANG, L. et al. (2009). SEER cancer statistics review, 1975–2006. National Cancer Institute, Bethesda, MD.

HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4** 251–257.

IVERSEN, E. S. JR. and CHEN, S. (2005). Population-calibrated gene characterization: Estimating age at onset distributions associated with cancer genes. *J. Amer. Statist. Assoc.* **100** 399–409. MR2170463 https://doi.org/10.1198/016214505000000196

JANOCHA, K. and CZARNECKI, W. M. (2017). On loss functions for deep neural networks in classification. Preprint. Available at arXiv:1702.05659.

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

JOHN, E. M., HOPPER, J. L., BECK, J. C., KNIGHT, J. A., NEUHAUSEN, S. L., SENIE, R. T., ZIOGAS, A., ANDRULIS, I. L., ANTON-CULVER, H. et al. (2004). The Breast Cancer Family Registry: An infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res.* **6** R375.

KATKI, H. A. (2006). Effect of misreported family history on Mendelian mutation prediction models. *Biometrics* **62** 478–487. MR2236830 https://doi.org/10.1111/j.1541-0420.2005.00488.x

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23** 462–466. MR0050243 https://doi.org/10.1214/aoms/1177729392
KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.
KOKUER, M., NAGUIB, R. N., JANCOVIC, P., YOUNGHUSBAND, H. B. and GREEN, R. (2006). A comparison of multi-layer neural network and logistic regression in hereditary non-polyposis colorectal cancer risk assessment. In 2005 *IEEE Engineering in Medicine and Biology* 27*th Annual Conference* 2417–2420. IEEE.
KRAFT, P. and THOMAS, D. C. (2000). Bias and efficiency in family-based gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *Am. J. Hum. Genet.* **66** 1119–1131.
KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.
LE BIHAN, C., MOUTOU, C., BRUGIÈRES, L., FEUNTEUN, J. and BONAÏTI-PELLIÉ, C. (1995). ARCAD: A method for estimating age-dependent disease risk associated with mutation carrier status from family data. *Genet. Epidemiol.* **12** 13–25.
LECUN, Y., BOTTOU, L., BENGIO, Y., HAFFNER, P. et al. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.
LESHNO, M., LIN, V. Y., PINKUS, A. and SCHOCKEN, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6** 861–867.
LI, Y., ZHANG, T., LIU, Z. and HU, H. (2017). A concatenating framework of shortcut convolutional neural networks. Preprint. Available at arXiv:1710.00974.
LI, O., LIU, H., CHEN, C. and RUDIN, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**. 1.
LITTLE, R. J. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* **793**. Wiley, New York.
MANDEL, J. C., KREDA, D. A., MANDL, K. D., KOHANE, I. S. and RAMONI, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *J. Am. Med. Inform. Assoc.* **23** 899–908. https://doi.org/10.1093/jamia/ocv189
MATSUNO, R. K., COSTANTINO, J. P., ZIEGLER, R. G., ANDERSON, G. L., LI, H., PEE, D. and GAIL, M. H. (2011). Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. *J. Natl. Cancer Inst.* **103** 951–961.
MCCARTHY, A. M., GUAN, Z., WELCH, M., GRIFFIN, M. E., SIPPO, D. A., DENG, Z., COOPEY, S. B., ACAR, A., SEMINE, A. et al. (2019). Performance of breast cancer risk assessment models in a large mammography cohort. *J. Natl. Cancer Inst.*
MIKI, Y., SWENSEN, J., SHATTUCK-EIDENS, D., FUTREAL, P. A., HARSHMAN, K., TAVTIGIAN, S., LIU, Q., COCHRAN, C., BENNETT, L. M. et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 66–71.
NEWCOMB, P. A., BARON, J., COTTERCHIO, M., GALLINGER, S., GROVE, J., HAILE, R., HALL, D., HOPPER, J. L., JASS, J. et al. (2007). Colon Cancer Family Registry: An international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomark. Prev.* **16** 2331–2343.
NIELSEN, M. A. (2015). *Neural Networks and Deep Learning* **25**. Determination Press, San Francisco, CA.
NIEPERT, M., AHMED, M. and KUTZKOV, K. (2016). Learning convolutional neural networks for graphs. In *International Conference on Machine Learning* 2014–2023.
OH, K.-S. and JUNG, K. (2004). GPU implementation of neural networks. *Pattern Recognit.* **37** 1311–1314.
COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2001). Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet* **358** 1389–1399.
PARMIGIANI, G., BERRY, D. A. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* **62** 145–158.
PATRO, S. and SAHU, K. K. (2015). Normalization: A preprocessing stage. Preprint. Available at arXiv:1503.06462.
PETERSEN, G. M., DE ANDRADE, M., GOGGINS, M., HRUBAN, R. H., BONDY, M., KORCZAK, J. F., GALLINGER, S., LYNCH, H. T., SYNGAL, S. et al. (2006). Pancreatic cancer genetic epidemiology consortium. *Cancer Epidemiol. Biomark. Prev.* **15** 704–710.
PICHERT, G., BOLLIGER, B., BUSER, K., PAGANI, O. and SWISS INSTITUTE FOR APPLIED CANCER RESEARCH NETWORK FOR CANCER PREDISPOSITION TESTING (2003). Evidence-based management options for women at increased breast/ovarian cancer risk. *Ann. Oncol.* **14** 9–19. https://doi.org/10.1093/annonc/mdg030
PORTNOI, T., YALA, A., SCHUSTER, T., BARZILAY, R., DONTCHOS, B., LAMB, L. and LEHMAN, C. (2019). Deep learning model to assess cancer risk on the basis of a breast MR image alone. *Am. J. Roentgenol.* **213** 227–233. https://doi.org/10.2214/AJR.18.20813

QUANTE, A. S., WHITTEMORE, A. S., SHRIVER, T., STRAUCH, K. and TERRY, M. B. (2012). Breast cancer risk assessment across the risk continuum: Genetic and nongenetic risk factors contributing to differential model performance. *Breast Cancer Res.* **14** R144. https://doi.org/10.1186/bcr3352

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

SHRIKUMAR, A., GREENSIDE, P. and KUNDAJE, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning* 3145–3153. PMLR.

SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958. MR3231592

STEFANSDOTTIR, V., JOHANNSSON, O. T., SKIRTON, H., TRYGGVADOTTIR, L., TULINIUS, H. and JONSSON, J. J. (2013). The use of genealogy databases for risk assessment in genetic health service: A systematic review. *J. Community Genet.* **4** 1–7. https://doi.org/10.1007/s12687-012-0103-3

STEFANSDOTTIR, V., SKIRTON, H., JOHANNSSON, O. T., OLAFSDOTTIR, H., OLAFSDOTTIR, G. H., TRYGGVADOTTIR, L. and JONSSON, J. J. (2019). Electronically ascertained extended pedigrees in breast cancer genetic counseling. *Fam. Cancer* **18** 153–160. https://doi.org/10.1007/s10689-018-0105-3

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

SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8**.

SWERDLOW, A. J., JONES, M. E., SCHOEMAKER, M. J., HEMMING, J., THOMAS, D., WILLIAMSON, J. and ASHWORTH, A. (2011). The breakthrough generations study: Design of a long-term UK cohort study to investigate breast cancer aetiology. *Br. J. Cancer* **105** 911–917.

TEAM, T. T. D., AL-RFOU, R., ALAIN, G., ALMAHAIRI, A., ANGERMUELLER, C., BAHDANAU, D., BALLAS, N., BASTIEN, F., BAYER, J. et al. (2016). Theano: A Python framework for fast computation of mathematical expressions. Preprint. Available at arXiv:1605.02688.

TEERLINK, C. C., ALBRIGHT, F. S., LINS, L. and CANNON-ALBRIGHT, L. A. (2012). A comprehensive survey of cancer risks in extended families. *Genet. Med.* **14** 107–114.

TERRY, M. B., LIAO, Y., WHITTEMORE, A. S., LEOCE, N., BUCHSBAUM, R., ZEINOMAR, N., DITE, G. S., CHUNG, W. K., KNIGHT, J. A. et al. (2019). 10-year performance of four models of breast cancer risk: A validation study. *Lancet Oncol.* **20** 504–517.

TICE, J. A., CUMMINGS, S. R., SMITH-BINDMAN, R., ICHIKAWA, L., BARLOW, W. E. and KERLIKOWSKE, K. (2008). Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Ann. Intern. Med.* **148** 337–347.

TRIPPA, L., WALDRON, L., HUTTENHOWER, C. and PARMIGIANI, G. (2015). Bayesian nonparametric cross-study validation of prediction methods. *Ann. Appl. Stat.* **9** 402–428. MR3341121 https://doi.org/10.1214/14-AOAS798

TYRER, J., DUFFY, S. W. and CUZICK, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23** 1111–1130.

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. MR2370850 https://doi.org/10.1198/016214507000000149

WANG, W., CHEN, S., BRUNE, K. A., HRUBAN, R. H., PARMIGIANI, G. and KLEIN, A. P. (2007). PancPRO: Risk assessment for individuals with a family history of pancreatic cancer. *J. Clin. Oncol.* **25** 1417–1422.

WANG, W., NIENDORF, K. B., PATEL, D., BLACKFORD, A., MARRONI, F., SOBER, A. J., PARMIGIANI, G. and TSAO, H. (2010). Estimating CDKN2A carrier probability and personalizing cancer risk assessments in hereditary melanoma using MelaPRO. *Cancer Res.* **70** 552–559.

WELCH, B. M., WILEY, K., PFLIEGER, L., ACHIANGIA, R., BAKER, K., HUGHES-HALBERT, C., MORRISON, H., SCHIFFMAN, J. and DOERR, M. (2018). Review and comparison of electronic patient-facing family health history tools. *J. Genet. Couns.* **27** 381–391. https://doi.org/10.1007/s10897-018-0235-7

WOOSTER, R., BIGNELL, G., LANCASTER, J., SWIFT, S., SEAL, S., MANGION, J., COLLINS, N., GREGORY, S., GUMBS, C. et al. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378** 789–792.

WU, Z., PAN, S., CHEN, F., LONG, G., ZHANG, C. and YU, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32** 4–24. MR4205495

YALA, A., LEHMAN, C., SCHUSTER, T., PORTNOI, T. and BARZILAY, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292** 60–66. https://doi.org/10.1148/radiol.2019182716

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

ZHANG, Q., NIAN WU, Y. and ZHU, S.-C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 8827–8836.

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

ZIOGAS, A. and ANTON-CULVER, H. (2003). Validation of family history data in cancer family registries. *Am. J. Prev. Med.* **24** 190–198.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 521–536
https://doi.org/10.1214/21-AOAS1511

IDENTIFYING INTERGENERATIONAL PATTERNS OF CORRELATED METHYLATION SITES

BY XICHEN MOU[1,a], HONGMEI ZHANG[1,b] AND S. HASAN ARSHAD[2,3,c]

[1]*Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis,* [a]*xmou@memphis.edu,* [b]*hzhang6@memphis.edu*

[2]*Allergy and Clinical Immunology, University of Southampton,* [c]*s.h.arshad@soton.ac.uk*

[3]*The David Hide Asthma and Allergy Research Centre, Isle of Wight*

DNA methylation can be transmitted through generations. This paper proposes a clustering method to identify the intergenerational patterns from parents to their offspring. Motivated by the potential of correlation between DNA methylation sites, we use the multivariate generalized beta distribution to model the blockwise correlation structure among the sites. A stochastic EM algorithm is implemented to estimate the parameters, and BIC is applied to determine the optimal number of clusters. Simulations demonstrate the feasibility of the proposed method. We further applied the approach to cluster DNA methylation data generated from a cohort study on asthma and allergic conditions.

REFERENCES

ARSHAD, S. H., KARMAUS, W., RAZA, A., KURUKULAARATCHY, R. J., MATTHEWS, S. M., HOLLOWAY, J. W., SADEGHNEJAD, A., ZHANG, H., ROBERTS, G. et al. (2012). The effect of parental allergy on childhood allergic diseases depends on the sex of the child. *J. Allergy Clin. Immunol.* **130** 427–434.

ARSHAD, S. H., HOLLOWAY, J. W., KARMAUS, W., ZHANG, H., EWART, S., MANSFIELD, L., MATTHEWS, S., HODGEKISS, C., ROBERTS, G. et al. (2018). Cohort profile: The Isle of Wight whole population birth cohort (IOWBC). *Int. J. Epidemiol.* **47** 1043–1044i.

ARYEE, M. J., JAFFE, A. E., CORRADA-BRAVO, H., LADD-ACOSTA, C., FEINBERG, A. P., HANSEN, K. D. and IRIZARRY, R. A. (2014). Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* **30** 1363–1369.

BELL, J. T., PAI, A. A., PICKRELL, J. K., GAFFNEY, D. J., PIQUE-REGI, R., DEGNER, J. F., GILAD, Y. and PRITCHARD, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12** R10. https://doi.org/10.1186/gb-2011-12-1-r10

BIBIKOVA, M., BARNES, B., TSAN, C., HO, V., KLOTZLE, B., LE, J. M., DELANO, D., ZHANG, L., SCHROTH, G. P. et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics* **98** 288–295.

CLARKE, R., RESSOM, H. W., WANG, A., XUAN, J., LIU, M. C., GEHAN, E. A. and WANG, Y. (2008). The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **8** 37–49.

CURLEY, J. P., MASHOODH, R. and CHAMPAGNE, F. A. (2011). Epigenetics and the origins of paternal effects. *Horm. Behav.* **59** 306–314.

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

ECKHARDT, F., LEWIN, J., CORTESE, R., RAKYAN, V. K., ATTWOOD, J., BURGER, M., BURTON, J., COX, T. V., DAVIES, R. et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38** 1378–1385.

HAN, S., ZHANG, H., LOCKETT, G. A., MUKHERJEE, N., HOLLOWAY, J. W. and KARMAUS, W. (2015). Identifying heterogeneous transgenerational DNA methylation sites via clustering in beta regression. *Ann. Appl. Stat.* **9** 2052–2072. MR3456365 https://doi.org/10.1214/15-AOAS865

HOFMEISTER, B. T., LEE, K., ROHR, N. A., HALL, D. W. and SCHMITZ, R. J. (2017). Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol.* **18** 1–16.

Key words and phrases. DNA methylation, EM, clustering, generalized beta distribution.

HOUSEMAN, E. A., CHRISTENSEN, B. C., YEH, R.-F., MARSIT, C. J., KARAGAS, M. R., WRENSCH, M., NELSON, H. H., WIEMELS, J., ZHENG, S. et al. (2008). Model-based clustering of DNA methylation array data: A recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinform.* **9** 365.

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

KAPPELER, L. and MEANEY, M. J. (2010). Epigenetics and parental effects. *BioEssays* **32** 818–827.

LEE, S.-H., PARK, J.-S. and PARK, C.-S. (2011). The search for genetic variants and epigenetics related to asthma. *Allergy, Asthma & Immunology Research* **3** 236–244.

LIBBY, D. L. and NOVICK, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *J. Educ. Stat.* **7** 271–294.

LOCKETT, G. A., PATIL, V. K., SOTO-RAMÍREZ, N., ZIYAB, A. H., HOLLOWAY, J. W. and KARMAUS, W. (2013). Epigenomics and allergic disease. *Epigenomics* **5** 685–699.

MOU, X., ZHANG, H. and ARSHAD, S. H. (2022a). Supplement to "Identifying intergenerational patterns of correlated methylation sites." https://doi.org/10.1214/21-AOAS1511SUPPA

MOU, X., ZHANG, H. and ARSHAD, S. H. (2022b). Computer codes for "Identifying intergenerational patterns of correlated methylation sites." https://doi.org/10.1214/21-AOAS1511SUPPB

NIELSEN, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli* **6** 457–489. MR1762556 https://doi.org/10.2307/3318671

PADMANABHAN, N., JIA, D., GEARY-JOO, C., WU, X., FERGUSON-SMITH, A. C., FUNG, E., BIEDA, M. C., SNYDER, F. F., GRAVEL, R. A. et al. (2013). Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell* **155** 81–93.

PARK, H.-S. and JUN, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36** 3336–3341.

QIN, L.-X. and SELF, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62** 526–533. MR2236835 https://doi.org/10.1111/j.1541-0420.2005.00498.x

SOUBRY, A., HOYO, C., JIRTLE, R. L. and MURPHY, S. K. (2014). A paternal environmental legacy: Evidence for epigenetic inheritance through the male germ line. *BioEssays* **36** 359–371.

STENZ, L., SCHECHTER, D. S., SERPA, S. R. and PAOLONI-GIACOBINO, A. (2018). Intergenerational transmission of DNA methylation signatures associated with early life stress. *Curr. Genomics* **19** 665–675.

WANG, D., YAN, L., HU, Q., SUCHESTON, L. E., HIGGINS, M. J., AMBROSONE, C. B., JOHNSON, C. S., SMIRAGLIA, D. J. and LIU, S. (2012). IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28** 729–730.

YU, F., XU, C., DENG, H.-W. and SHEN, H. (2020). A novel computational strategy for DNA methylation imputation using mixture regression model (MRM). *BMC Bioinform.* **21** 1–17.

ZHANG, W., SPECTOR, T. D., DELOUKAS, P., BELL, J. T. and ENGELHARDT, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16** 14. https://doi.org/10.1186/s13059-015-0581-9

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 537–550
https://doi.org/10.1214/21-AOAS1513

ORDINAL PROBIT FUNCTIONAL OUTCOME REGRESSION WITH APPLICATION TO COMPUTER-USE BEHAVIOR IN RHESUS MONKEYS

BY MARK J. MEYER[1,a], JEFFREY S. MORRIS[2,b], REGINA PAXTON GAZES[3,c] AND BRENT A. COULL[4,d]

[1]*Department of Mathematics and Statistics, Georgetown University,* [a]*mjm556@georgetown.edu*
[2]*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania,* [b]*Jeffrey.Morris@pennmedicine.upenn*
[3]*Department of Psychology and Program in Animal Behavior, Bucknell University,* [c]*reggie.gazes@bucknell.edu*
[4]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [d]*bcoull@hsph.harvard.edu*

Research in functional regression has made great strides in expanding to non-Gaussian functional outcomes, but exploration of ordinal functional outcomes remains limited. Motivated by a study of computer-use behavior in rhesus macaques (*Macaca mulatta*), we introduce the ordinal probit functional outcome regression model (OPFOR). OPFOR models can be fit using one of several basis functions including penalized B-splines, wavelets, and O'Sullivan splines—the last of which typically performs best. Simulation using a variety of underlying covariance patterns shows that the model performs reasonably well in estimation under multiple basis functions with near nominal coverage for joint credible intervals. Finally, in application we use Bayesian model selection criteria adapted to functional outcome regression to best characterize the relation between several demographic factors of interest and the monkeys' computer use over the course of a year. In comparison with a standard ordinal longitudinal analysis, OPFOR outperforms a cumulative-link mixed-effects model in simulation and provides additional and more nuanced information on the nature of the monkeys' computer-use behavior.

REFERENCES

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

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

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

CHRISTENSEN, R. H. B. (2019). ordinal—Regression Models for Ordinal Data. R package version 2019.4-25. http://www.cran.r-project.org/package=ordinal/.

DE WAAL, D. J. (2014). Matrix-valued distributions. In *Wiley StatsRef: Statistics Reference Online* (N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels, eds.). https://doi.org/10.1002/9781118445112.stat01061

FAGOT, J. and BONTÉ, E. (2010). Automated testing of cognitive performance in monkeys: Use of a battery of computerized test systems by a troop of semi-free-ranging baboons (Papio papio). *Behav. Res. Methods* **42** 507–516. https://doi.org/10.3758/BRM.42.2.507

FARAWAY, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39** 254–261. MR1462586 https://doi.org/10.2307/1271130

GAZES, R. P., BROWN, E. K., BASILE, B. M. and HAMPTON, R. R. (2013). Automated cognitive testing of monkeys in social groups yields results comparable to individual laboratory based testing. *Anim. Cogn.* **16** 445–458.

GAZES, R. P., LUTZ, M. D., MEYER, M. J., HASSETT, T. and HAMPTON, R. R. (2019). Influences of demographic, seasonal, and social factors on automated touchscreen computer use by a socially-housed group of rhesus monkeys (Macaca mulatta). *PLoS ONE* **14** e0215060.

Key words and phrases. Functional data analysis, ordinal variates, O'Sullivan splines, wavelets, automated cognitive testing, probit regression.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. CRC Press–CRC, Boca Raton, FL.

GERTHEISS, J., MAIER, V., HESSEL, E. F. and STAICU, A.-M. (2015). Marginal functional regression models for analyzing the feeding behavior of pigs. *J. Agric. Biol. Environ. Stat.* **20** 353–370. MR3396556 https://doi.org/10.1007/s13253-015-0212-7

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

GOLDSMITH, J. and SCHWARTZ, J. E. (2017). Variable selection in the functional linear concurrent model. *Stat. Med.* **36** 2237–2250. MR3660128 https://doi.org/10.1002/sim.725

GOLDSMITH, J., ZIPUNNIKOV, V. and SCHRACK, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **71** 344–353. MR3366239 https://doi.org/10.1111/biom.12278

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x

HALL, P., MÜLLER, H.-G. and YAO, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 703–723. MR2523900 https://doi.org/10.1111/j.1467-9868.2008.00656.x

KRAFTY, R. T., GIMOTTY, P. A., HOLTZ, D., COUKOS, G. and GUO, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics* **64** 1023–1031. MR2522249 https://doi.org/10.1111/j.1541-0420.2007.00980.x

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

LI, G., HUANG, J. Z. and SHEN, H. (2018). Exponential family functional data analysis via a low-rank model. *Biometrics* **74** 1301–1310. MR3908148 https://doi.org/10.1111/biom.12885

LI, H., STAUDENMAYER, J. and CARROLL, R. J. (2014). Hierarchical functional data with mixed continuous and binary measurements. *Biometrics* **70** 802–811. MR3295741 https://doi.org/10.1111/biom.12211

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

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

MEYER, M. J., MORRIS, J. S., GAZES, R. P. and COULL, B. A. (2022). Supplement to "Ordinal Probit Functional Outcome Regression with Application to Computer-Use Behavior in Rhesus Monkeys." https://doi.org/10.1214/21-AOAS1513SUPPA, https://doi.org/10.1214/21-AOAS1513SUPPB

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

RAMSAY, J. O. and SILVERMAN, B. W. (1997). *Functional Data Analysis*, 1st ed. Springer, New York.

REISS, P. T., HUANG, L. and MENNES, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *Int. J. Biostat.* **6** 28. MR2683940 https://doi.org/10.2202/1557-4679.1246

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

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

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

SHI, J. Q., WANG, B., MURRAY-SMITH, R. and TITTERINGTON, D. M. (2007). Gaussian process functional regression modeling for batch data. *Biometrics* **63** 714–723. MR2395708 https://doi.org/10.1111/j.1541-0420.2007.00758.x

VAN DER LINDE, A. (2009). A Bayesian latent variable approach to functional principal components analysis with binary and count data. *AStA Adv. Stat. Anal.* **93** 307–333. MR2545698 https://doi.org/10.1007/s10182-009-0113-6

VAN DER LINDE, A. (2011). Reduced rank regression models with latent variables in Bayesian functional data analysis. *Bayesian Anal.* **6** 77–126. MR2781809 https://doi.org/10.1214/11-BA603

WAND, M. P. and ORMEROD, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Aust. N. Z. J. Stat.* **50** 179–198. MR2431193 https://doi.org/10.1111/j.1467-842X.2008.00507.x

WANG, B. and SHI, J. Q. (2014). Generalized Gaussian process regression model for non-Gaussian functional data. *J. Amer. Statist. Assoc.* **109** 1123–1133. MR3265685 https://doi.org/10.1080/01621459.2014.889021

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2011). Robust, adaptive functional regression in functional mixed model framework. *J. Amer. Statist. Assoc.* **106** 1167–1179. MR2894772 https://doi.org/10.1198/jasa.2011.tm10370

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260–1268. MR3040032 https://doi.org/10.1111/j.1541-0420.2012.01765.x

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 551–572
https://doi.org/10.1214/21-AOAS1516

PARTITIONING AROUND MEDOIDS CLUSTERING AND RANDOM FOREST CLASSIFICATION FOR GIS-INFORMED IMPUTATION OF FLUORIDE CONCENTRATION DATA

BY YU GU[1,a], JOHN S. PREISSER[1,b], DONGLIN ZENG[1,c], POOJAN SHRESTHA[2,3,g], MOLINA SHAH[2,d], MIGUEL A. SIMANCAS-PALLARES[2,e], JEANNIE GINNIS[2,f] AND KIMON DIVARIS[2,3,h]

[1]*Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill,* [a]*yugu@live.unc.edu,* [b]*jpreisse@bios.unc.edu,* [c]*dzeng@email.unc.edu*

[2]*Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill,* [d]*molina.shah@unc.edu,* [e]*simancas@email.unc.edu,* [f]*jeannie_ginnis@unc.edu*

[3]*Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill,* [g]*poojansh@live.unc.edu,* [h]*Kimon_Divaris@unc.edu*

Community water fluoridation is an important component of oral health promotion, as fluoride exposure is a well-documented dental caries-preventive agent. Direct measurements of domestic water fluoride content provide valuable information regarding individuals' fluoride exposure and thus caries risk; however, they are logistically challenging to carry out at a large scale in oral health research. This article describes the development and evaluation of a novel method for the imputation of missing domestic water fluoride concentration data informed by spatial autocorrelation. The context is a state-wide epidemiologic study of pediatric oral health in North Carolina, where domestic water fluoride concentration information was missing for approximately 75% of study participants with clinical data on dental caries. A new machine-learning-based imputation method that combines partitioning around medoids clustering and random forest classification (PAMRF) is developed and implemented. Imputed values are filtered according to allowable error rates or target sample size, depending on the requirements of each application. In leave-one-out cross-validation and simulation studies, PAMRF outperforms four existing imputation approaches—two conventional spatial interpolation methods (i.e., inverse-distance weighting, IDW and universal kriging, UK) and two supervised learning methods (k-nearest neighbors, KNN, and classification and regression trees, CART). The inclusion of multiply imputed values in the estimation of the association between fluoride concentration and dental caries prevalence resulted in essentially no change in PAMRF estimates but substantial gains in precision due to larger effective sample size. PAMRF is a powerful new method for the imputation of missing fluoride values where geographical information exists.

REFERENCES

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

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

BREIMAN, L. (2002). Manual on Setting up, Using, and Understanding Random Forests v3. 1. Statistics Department University of California Berkeley, CA, USA **1** 58.

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

BRUNELLE, J. and CARLOS, J. (1990). Recent trends in dental caries in US children and the effect of water fluoridation. *J. Dent. Res.* **69** 723–727.

BUUREN, S. V. and GROOTHUIS-OUDSHOORN, K. (2010). Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 1–68.

Key words and phrases. Missing values, spatial interpolation, clustering, multiple imputation, random forest.

CATE, J. M. T. (1999). Current concepts on the theories of the mechanism of action of fluoride. *Acta Odontol. Scand.* **57** 325–329.
CHEN, J. and SHAO, J. (2000). Nearest neighbor imputation for survey data. *J. Off. Stat.* **16** 113.
CLIFF, A. D. and ORD, J. K. (1981). *Spatial Processes*: *Models & Applications*. Pion Ltd., London. MR0632256
CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1239641 https://doi.org/10.1002/9781119115151
DIVARIS, K. and JOSHI, A. (2020). The building blocks of precision oral health in early childhood: The ZOE 2.0 study. *J. Public Health Dent.* **80** S31–S36. https://doi.org/10.1111/jphd.12303
DIVARIS, K., SLADE, G. D., FERREIRA ZANDONA, A. G., PREISSER, J. S., GINNIS, J., SIMANCAS-PALLARES, M. A., AGLER, C. S., SHRESTHA, P., KARHADE, D. S. et al. (2020). Cohort profile: ZOE 2.0—a community-based genetic epidemiologic study of early childhood oral health. *Int. J. Environ. Res. Public Health* **17** 8056.
ECKERT, S., FEINGOLD, E., COOPER, M., VANYUKOV, M. M., MAHER, B. S., SLAYTON, R. L., WILLING, M. C., REIS, S. E., MCNEIL, D. W. et al. (2017). Variants on chromosome 4q21 near PKD2 and SIBLINGs are associated with dental caries. *J. Hum. Genet.* **62** 491–496.
FALKOWSKI, M. J., HUDAK, A. T., CROOKSTON, N. L., GESSLER, P. E., UEBLER, E. H. and SMITH, A. M. (2010). Landscape-scale parameterization of a tree-level forest growth model: A K-nearest neighbor imputation approach incorporating LiDAR data. *Can. J. For. Res.* **40** 184–199.
FISHER-OWENS, S. A., GANSKY, S. A., PLATT, L. J., WEINTRAUB, J. A., SOOBADER, M.-J., BRAMLETT, M. D. and NEWACHECK, P. W. (2007). Influences on children's oral health: A conceptual model. *Pediatrics* **120** e510–e520.
FRANKE, R. (1982). Scattered data interpolation: Tests of some methods. *Math. Comp.* **38** 181–200. MR0637296 https://doi.org/10.2307/2007474
GINNIS, J., ZANDONÁ, A. G. F., SLADE, G. D., CANTRELL, J., ANTONIO, M. E., PAHEL, B. T., MEYER, B. D., SHRESTHA, P., SIMANCAS-PALLARES, M. A. et al. (2019). Measurement of early childhood oral health for research purposes: Dental caries experience and developmental defects of the enamel in the primary dentition. In *Odontogenesis* 511–523. Springer, Berlin.
GOWER, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27** 857–871.
GU, Y., PREISSER, J. S., ZENG, D., SHRESTHA, P., SHAH, M., SIMANCAS-PALLARES, M. A., GINNIS, J. and DIVARIS, K. (2022). Supplement to "Partitioning around medoids clustering and random forest classification for GIS-informed imputation of fluoride concentration data." https://doi.org/10.1214/21-AOAS1516SUPP
HA, D. H., SPENCER, A. J., PERES, K. G., RUGG-GUNN, A. J., SCOTT, J. A. and DO, L. G. (2019). Fluoridated water modifies the effect of breastfeeding on dental caries. *J. Dent. Res.* **98** 755–762. https://doi.org/10.1177/0022034519843487
HASTIE, T., TIBSHIRANI, R., SHERLOCK, G., EISEN, M., BROWN, P. and BOTSTEIN, D. (1999). Imputing Missing Data for Gene Expression Arrays. Stanford University Statistics Department Technical Report.
HENNIG, C. and LIAO, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification (with discussion). *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 309–369. MR3060621 https://doi.org/10.1111/j.1467-9876.2012.01066.x
IHEOZOR-EJIOFOR, Z., WORTHINGTON, H. V., WALSH, T., O'MALLEY, L., CLARKSON, J. E., MACEY, R., ALAM, R., TUGWELL, P., WELCH, V. et al. (2015). Water fluoridation for the prevention of dental caries. *Cochrane Database Syst. Rev.* **6**.
JOHNSTON, K., VER HOEF, J. M., KRIVORUCHKO, K. and LUCAS, N. (2001). *Using ArcGIS Geostatistical Analyst* **380**. Esri, Redlands.
KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding Groups in Data*: *An Introduction to Cluster Analysis* **344**. Wiley, New York.
LAM, N. S.-N. (1983). Spatial interpolation methods: A review. *Am. Cartogr.* **10** 129–150.
LIAW, A., WIENER, M. et al. (2002). Classification and regression by randomForest. *R News* **2** 18–22.
MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M. and HORNIK, K. (2019). cluster: Cluster Analysis Basics and Extensions. R Package Version 2.1. 0. 2019.
MITAS, L. and MITASOVA, H. (1999). Spatial interpolation. In *Geographical Information Systems*: *Principles*, *Techniques*, *Management and Applications* (P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. E. . Rhind, eds.) **1** 481–492 34. Wiley, New York.
MORAN, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37** 17–23. MR0035933 https://doi.org/10.1093/biomet/37.1-2.17
RCOLORBREWER, S. and LIAW, M. A. (2018). Package 'randomForest'. University of California, Berkeley: Berkeley, CA, USA.
ROUSSEEUW, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20** 53–65.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

SELWITZ, R. H., ISMAIL, A. I. and PITTS, N. B. (2007). Dental caries. *Lancet* **369** 51–59.

SHAFFER, J., WANG, X., FEINGOLD, E., LEE, M., BEGUM, F., WEEKS, D., CUENCO, K., BARMADA, M., WENDELL, S. et al. (2011). Genome-wide association scan for childhood caries implicates novel genes. *J. Dent. Res*. **90** 1457–1462.

SHEPARD, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the* 1968 23*rd ACM National Conference* 517–524. ACM, New York.

SU, L., TOM, B. D. and FAREWELL, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10** 374–389.

THERNEAU, T. and ATKINSON, B. (2008). rpart: Recursive Partitioning. R port by Brian Ripley. R Package Version 3–1.

VAN DE VELDEN, M., IODICE D'ENZA, A. and MARKOS, A. (2019). Distance-based clustering of mixed data. *Wiley Interdiscip. Rev.*: *Comput. Stat*. **11** e1456, 12. MR3950393 https://doi.org/10.1002/wics.1456

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 573–595
https://doi.org/10.1214/21-AOAS1517

BAYESIAN MITIGATION OF SPATIAL COARSENING FOR A HAWKES MODEL APPLIED TO GUNFIRE, WILDFIRE AND VIRAL CONTAGION

BY ANDREW J. HOLBROOK[1,a], XIANG JI[2,b] AND MARC A. SUCHARD[3,c]

[1]*Department of Biostatistics, UCLA,* [a]*aholbroo@g.ucla.edu*
[2]*Department of Mathematics, Tulane University,* [b]*xji4@tulane.edu*
[3]*Departments of Biostatistics, Human Genetics and Computational Medicine, UCLA,* [c]*msuchard@ucla.edu*

Self-exciting spatiotemporal Hawkes processes have found increasing use in the study of large-scale public health threats, ranging from gun violence and earthquakes to wildfires and viral contagion. Whereas many such applications feature locational uncertainty, that is, the exact spatial positions of individual events are unknown, most Hawkes model analyses to date have ignored spatial coarsening present in the data. Three particular 21st century public health crises—urban gun violence, rural wildfires and global viral spread—present qualitatively and quantitatively varying uncertainty regimes that exhibit: (a) different collective magnitudes of spatial coarsening, (b) uniform and mixed magnitude coarsening, (c) differently shaped uncertainty regions and—less orthodox—(d) locational data distributed within the "wrong" effective space. We explicitly model such uncertainties in a Bayesian manner and jointly infer unknown locations together with all parameters of a reasonably flexible Hawkes model, obtaining results that are practically and statistically distinct from those obtained while ignoring spatial coarsening. This work also features two different secondary contributions: first, to facilitate Bayesian inference of locations and background rate parameters, we make a subtle yet crucial change to an established kernel-based rate model, and second, to facilitate the same Bayesian inference at scale, we develop a massively parallel implementation of the model's log-likelihood gradient with respect to locations and thus avoid its quadratic computational cost in the context of Hamiltonian Monte Carlo. Our examples involve thousands of observations and allow us to demonstrate practicality at moderate scales.

REFERENCES

BEDFORD, T., SUCHARD, M. A., LEMEY, P., DUDAS, G., GREGORY, V., HAY, A. J., MCCAULEY, J. W., RUSSELL, C. A., SMITH, D. J. et al. (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3** e01914. https://doi.org/10.7554/eLife.01914

BEDFORD, T., RILEY, S., BARR, I. G., BROOR, S., CHADHA, M., COX, N. J., DANIELS, R. S., GUNASEKARAN, C. P., HURT, A. C. et al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523** 217–220. PMCID: PMC4499780.

BROCKMANN, D. and HELBING, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science* **342** 1337–1342.

CARR, J. and DOLEAC, J. L. (2016). The geography, incidence, and underreporting of gun violence: New evidence using ShotSpotter data. *Brookings* (April 26, 2016).

CHOI, E., DU, N., CHEN, R., SONG, L. and SUN, J. (2015). Constructing disease network and temporal progression model via context-sensitive Hawkes process. In 2015 *IEEE International Conference on Data Mining* 721–726. IEEE, New York.

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

DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II: General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

Key words and phrases. Bayesian multidimensional scaling, gun violence, self-exciting processes, spatial coarsening, viral contagion, wildfires.

DeSarbo, W. S., Kim, Y. and Fong, D. (1999). A Bayesian multidimensional scaling procedure for the spatial analysis of revealed choice data. *J. Econometrics* **89** 79–108. MR1681137 https://doi.org/10.1016/S0304-4076(98)00056-6

Fox, E. W., Schoenberg, F. P. and Gordon, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.* **10** 1725–1756. MR3553242 https://doi.org/10.1214/16-AOAS957

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood.

Gilks, W. R., Best, N. G. and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **44** 455–472.

Hawkes, A. (1973). Cluster models for earthquakes-regional comparisons. *Bull. Int. Stat. Inst.* **45** 454–461.

Heitjan, D. F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* **49** 1099–1109.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174 https://doi.org/10.1214/aos/1176348396

Holbrook, A. J., Ji, X. and Suchard, M. A. (2022). Supplement to "Bayesian mitigation of spatial coarsening for a Hawkes model applied to gunfire, wildfire and viral contagion." https://doi.org/10.1214/21-AOAS1517SUPPA, https://doi.org/10.1214/21-AOAS1517SUPPB, https://doi.org/10.1214/21-AOAS1517SUPPC

Holbrook, A. J., Loeffler, C. E., Flaxman, S. R. and Suchard, M. A. (2021a). Scalable Bayesian inference for self-excitatory stochastic processes applied to big American gunfire data. *Stat. Comput.* **31** Paper No. 4, 15 pp. MR4199464 https://doi.org/10.1007/s11222-020-09980-4

Holbrook, A. J., Lemey, P., Baele, G., Dellicour, S., Brockmann, D., Rambaut, A. and Suchard, M. A. (2021b). Massive parallelization boosts big Bayesian multidimensional scaling. *J. Comput. Graph. Statist.* **30** 11–24. MR4235961 https://doi.org/10.1080/10618600.2020.1754226

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *R J.* **5** 144–161.

Kelly, J. D., Park, J., Harrigan, R. J., Hoff, N. A., Lee, S. D., Wannier, R., Selo, B., Mossoko, M., Njoloko, B. et al. (2019). Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. *Epidemics* **28** 100354.

Kim, H. (2011). Spatio-temporal point process models for the spread of avian influenza virus (H5N1). Ph.D. thesis, UC Berkeley. MR2926851

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29** 1–27. MR0169712 https://doi.org/10.1007/BF02289565

Loeffler, C. and Flaxman, S. (2018). Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.* **34** 999–1017.

Lomax, A., Michelini, A. and Curtis, A. (2009). Earthquake location, direct, global-search methods. *Encycl. Complex. Syst. Sci.* **5** 2449–2473.

Marquis de Laplace, P. S. (1825). *Essai philosophique sur les probabilités*. Bachelier.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8** 1612–1639. MR3271346 https://doi.org/10.1214/14-AOAS743

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.

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

Nishimura, A. and Suchard, M. A. (2018). Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in "large n & large p" sparse Bayesian regression. Preprint. Available at arXiv:1810.12437.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

Oh, M.-S. and Raftery, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *J. Amer. Statist. Assoc.* **96** 1031–1044. MR1947251 https://doi.org/10.1198/016214501753208690

Oh, M.-S. and Raftery, A. E. (2007). Model-based clustering with dissimilarities: A Bayesian approach. *J. Comput. Graph. Statist.* **16** 559–585. MR2351080 https://doi.org/10.1198/106186007X236127

Park, J., Chaffee, A. W., Harrigan, R. J. and Schoenberg, F. P. (2018). A non-parametric Hawkes model of the spread of Ebola in West Africa. *J. Appl. Stat.* To appear. https://doi.org/10.1080/02664763.2020.1825646

Park, J., Schoenberg, F. P., Bertozzi, A. L. and Brantingham, P. J. (2019). Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2** 559–572.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2014). The Bayesian bridge. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 713–733. MR3248673 https://doi.org/10.1111/rssb.12042

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

RAMSAY, J. O. (1982). Some statistical approaches to multidimensional scaling data. *J. Roy. Statist. Soc. Ser. A* **145** 285–312. MR0678529 https://doi.org/10.2307/2981865

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

RIZOIU, M.-A., MISHRA, S., KONG, Q., CARMAN, M. and XIE, L. (2018). SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the* 2018 *World Wide Web Conference on World Wide Web* 419–428. International World Wide Web Conferences Steering Committee.

ROBERTS, G. O. and ROSENTHAL, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* **44** 458–475. MR2340211 https://doi.org/10.1239/jap/1183667414

SCHOENBERG, F. P. (2004). Testing separability in spatial-temporal marked point processes. *Biometrics* **60** 471–481. MR2066282 https://doi.org/10.1111/j.0006-341X.2004.00192.x

SCHOENBERG, F. P. (2013). Facilitated estimation of ETAS. *Bull. Seismol. Soc. Amer.* **103** 601–605.

SCHOENBERG, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statist. Sinica* **26** 861–879. MR3497774

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4** vey016. https://doi.org/10.1093/ve/vey016

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

WICKHAM, H. (2016). *Ggplot2*: *Elegant Graphics for Data Analysis*. Springer, New York.

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res., Solid Earth* **109** B05301.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 596–611
https://doi.org/10.1214/21-AOAS1518

ACCOUNTING FOR DROP-OUT USING INVERSE PROBABILITY CENSORING WEIGHTS IN LONGITUDINAL CLUSTERED DATA WITH INFORMATIVE CLUSTER SIZE

BY AYA A. MITANI[1,a], ELIZABETH K. KAYE[2,b] AND KERRIE P. NELSON[3,c]

[1]*Division of Biostatistics, University of Toronto Dalla Lana School of Public Health,* [a]*aya.mitani@utoronto.ca*
[2]*Department of Health Policy and Health Services Research, Boston University Henry M. Goldman School of Dental Medicine,* [b]*kralle@bu.edu*
[3]*Department of Biostatistics, Boston University School of Public Health,* [c]*kerrie@bu.edu*

Periodontal disease is a serious gum infection impacting half of the U.S. adult population that may lead to loss of teeth. Using standard marginal models to study the association between patient-level predictors and tooth-level outcomes can lead to biased estimates because the independence assumption between the outcome (periodontal disease) and cluster size (number of teeth per patient) is violated. Specifically, the baseline number of teeth of a patient is informative. In this setting a cluster-weighted generalized estimating equations (CWGEE) approach can be used to obtain unbiased marginal inference from data with informative cluster size (ICS). However, in many longitudinal studies of dental health, including the Veterans Affairs Dental Longitudinal Study, the rate of tooth-loss or tooth drop-out over time is also informative, creating a missing at random data mechanism. Here, we propose a novel modeling approach that incorporates the technique of inverse probability censoring weights into CWGEE with binary outcomes to account for ICS and informative drop-out over time. In an extensive simulation study we demonstrate that results obtained from our proposed method yield lower bias and excellent coverage probability, compared to those obtained from traditional methods which do not account for ICS or drop-out.

REFERENCES

BENHIN, E., RAO, J. N. K. and SCOTT, A. J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92** 435–450. MR2201369 https://doi.org/10.1093/biomet/92.2.435

BIBLE, J., BECK, J. D. and DATTA, S. (2016). Cluster adjusted regression for displaced subject data (CARDS): Marginal inference under potentially informative temporal cluster size profiles. *Biometrics* **72** 441–451. MR3515771 https://doi.org/10.1111/biom.12456

CHAGANTY, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plann. Inference* **63** 39–54. MR1474184 https://doi.org/10.1016/S0378-3758(96)00203-0

FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2004). *Applied Longitudinal Analysis*. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2063401

HOFFMAN, E. B., SEN, P. K. and WEINBERG, C. R. (2001). Within-cluster resampling. *Biometrika* **88** 1121–1134. MR1872223 https://doi.org/10.1093/biomet/88.4.1121

HOGAN, J. W., ROY, J. and KORKONTZELOU, C. (2004). Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Stat. Med.* **23** 1455–1497.

KAPUR, K. K., GLASS, R. L., LOFTUS, E. R., ALMAN, J. E. and FELLER, R. P. (1972). The veterans administration longitudinal study of oral health and disease. *Aging Hum. Dev.* **3** 125–137. https://doi.org/10.2190/WLL4-ET76-UQWN-R5FL

KAYE, E. K., CHEN, N., CABRAL, H. J., VOKONAS, P. and GARCIA, R. I. (2016). Metabolic syndrome and periodontal disease progression in men. *J. Dent. Res.* **95** 822–828. https://doi.org/10.1177/0022034516641053

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

Key words and phrases. Clustered data, longitudinal data, generalized estimating equations, periodontal disease, tooth loss.

MITANI, A. A., KAYE, E. K. and NELSON, K. P. (2019). Marginal analysis of ordinal clustered longitudinal data with informative cluster size. *Biometrics* **75** 938–949. MR4012099 https://doi.org/10.1111/biom.13050

MITANI, A. A., KAYE, E. K. and NELSON, K. P. (2022). Supplement to "Accounting for drop-out using inverse probability censoring weights in longitudinal clustered data with informative cluster size." https://doi.org/10.1214/21-AOAS1518SUPPA, https://doi.org/10.1214/21-AOAS1518SUPPB

PARZEN, M., GHOSH, S., LIPSITZ, S., SINHA, D., FITZMAURICE, G. M., MALLICK, B. K. and IBRAHIM, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Ann. Appl. Stat.* **5** 449–467. MR2810405 https://doi.org/10.1214/10-AOAS390

PREISSER, J. S., LOHMAN, K. K. and RATHOUZ, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Stat. Med.* **21** 3035–3054.

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

WANG, M., KONG, M. and DATTA, S. (2011). Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Stat. Methods Med. Res.* **20** 347–367. MR2829115 https://doi.org/10.1177/0962280209347043

WANG, C. and PAIK, M. C. (2011). A weighting approach for GEE analysis with missing data. *Comm. Statist. Theory Methods* **40** 2397–2411. MR2863213 https://doi.org/10.1080/03610921003764282

WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471 https://doi.org/10.1111/1541-0420.00005

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 612–624
https://doi.org/10.1214/21-AOAS1520

LIKELIHOOD-BASED BACTERIAL IDENTIFICATION APPROACH FOR BIMICROBIAL MASS SPECTROMETRY DATA

BY SO YOUNG RYU[a]

School of Public Health, University of Nevada, Reno, [a]*soyoungr@unr.edu*

Mass spectrometry is a potential diagnostic tool for rapid bacterial detection. However, in order to use this technology in clinical settings, it is important to develop sound statistical algorithms that can accurately analyze polymicrobial mass spectrometry data. Here, we propose a likelihood-based bacterial identification algorithm for bimicrobial mass spectrometry data. Specifically, we introduce a two-component mixture model with partially known labels. This method can model peaks with unknown origins. It also considers errors in mass-to-charge ratios and intensities of peaks between observed and reference mass spectra. Coupled with a decoy strategy, the likelihood is used to identify bacterial species and to measure uncertainty of such identifications. Using two real mass spectrometry datasets, we demonstrate the superior performance of our approach in accurate bacterial identifications, compared to model-free approaches. Example datasets and R codes for the proposed method are freely available under MIT license at https://github.com/soyoungryu/BacID.

REFERENCES

APPALA, K., BIMPEH, K., FREEMAN, C. and HINES, K. M. (2020). Recent applications of mass spectrometry in bacterial lipidomics. *Anal. Bioanal. Chem.* **412** 5935–5943. https://doi.org/10.1007/s00216-020-02541-8

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

CHAPMAN, J. D., GOODLETT, D. R. and MASSELON, C. D. (2014). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* **33** 452–470.

CHOI, H. and NESVIZHSKII, A. I. (2008). Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **7** 254–265.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537

FEUCHEROLLES, M., POPPERT, S., UTZINGER, J. and BECKER, S. L. (2019). MALDI-TOF mass spectrometry as a diagnostic tool in human and veterinary helminthology: A systematic review. *Parasites Vectors* **12** 245. https://doi.org/10.1186/s13071-019-3493-9

FONDRIE, W. E., LIANG, T., OYLER, B. L., LEUNG, L. M., ERNST, R. K., STRICKLAND, D. K. and GOODLETT, D. R. (2018). Pathogen identification direct from polymicrobial specimens using membrane glycolipids. *Sci. Rep.* **8** 1–11.

GRAY, T. J., THOMAS, L., OLMA, T., IREDELL, J. R. and CHEN, S. C.-A. (2013). Rapid identification of Gram-negative organisms from blood culture bottles using a modified extraction method and MALDI-TOF mass spectrometry. *Diagn. Microbiol. Infect. Dis.* **77** 110–112.

KÄLL, L., CANTERBURY, J. D., WESTON, J., NOBLE, W. S. and MACCOSS, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4** 923–925.

KÄLL, L., STOREY, J. D., MACCOSS, M. J. and NOBLE, W. S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **7** 29–34. https://doi.org/10.1021/pr700600n

LAWSON, C. L. and HANSON, R. J. (1995). *Solving Least Squares Problems. Classics in Applied Mathematics* **15**. SIAM, Philadelphia, PA. Revised reprint of the 1974 original. MR1349828 https://doi.org/10.1137/1.9781611971217

Key words and phrases. Bacterial identification, mass spectrometry, likelihood-based scoring approach, mixture model.

LEUNG, L. M., FONDRIE, W. E., DOI, Y., JOHNSON, J. K., STRICKLAND, D. K., ERNST, R. K. and GOODLETT, D. R. (2017). Identification of the ESKAPE pathogens by mass spectrometric analysis of microbial membrane glycolipids. *Sci. Rep.* **7** 1–10.

LEWIS, N. H., HITCHCOCK, D. B., DRYDEN, I. L. and ROSE, J. R. (2018). Peptide refinement by using a stochastic search. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1207–1236. MR3873706 https://doi.org/10.1111/rssc.12280

LI, Q., ENG, J. K. and STEPHENS, M. (2012). A likelihood-based scoring method for peptide identification using mass spectrometry. *Ann. Appl. Stat.* **6** 1775–1794. MR3058683 https://doi.org/10.1214/12-AOAS568

LOURENS, S., ZHANG, Y., LONG, J. D. and PAULSEN, J. S. (2013). Bias in estimation of a mixture of normal distributions. *J. Biometr. Biostat.* **4**.

MAHÉ, P., ARSAC, M., CHATELLIER, S., MONNIN, V., PERROT, N., MAILLER, S., GIRARD, V., RAMJEET, M., SURRE, J. et al. (2014). Automatic identification of mixed bacterial species fingerprints in a MALDI-TOF mass-spectrum. *Bioinformatics* **30** 1280–1286.

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MÖRTELMAIER, C., PANDA, S., ROBERTSON, I., KRELL, M., CHRISTODOULOU, M., REICHARDT, N. and MULDER, I. (2019). Identification performance of MALDI-ToF-MS upon mono- and bi-microbial cultures is cell number and culture proportion dependent. *Anal. Bioanal. Chem.* **411** 7027–7038. https://doi.org/10.1007/s00216-019-02080-x

PANCHAUD, A., SCHERL, A., SHAFFER, S. A., VON HALLER, P. D., KULASEKARA, H. D., MILLER, S. I. and GOODLETT, D. R. (2009). Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Anal. Chem.* **81** 6481–6488.

RYU, S. Y. (2022a). Supplement to "Likelihood-based bacterial identification approach for bimicrobial mass spectrometry data." https://doi.org/10.1214/21-AOAS1520SUPPA

RYU, S. Y. (2022b). Supplement to "Likelihood-based bacterial identification approach for bimicrobial mass spectrometry data." https://doi.org/10.1214/21-AOAS1520SUPPB

RYU, S., GOODLETT, D. R., NOBLE, W. S. and MININ, V. N. (2012). A statistical approach to peptide identification from clustered tandem mass spectrometry data. In 2012 *IEEE International Conference on Bioinformatics and Biomedicine Workshops* 648–653. IEEE.

RYU, S. Y., WENDT, G. A., CHANDLER, C. E., ERNST, R. K. and GOODLETT, D. R. (2019). Model-based spectral library approach for bacterial identification via membrane glycolipids. *Anal. Chem.* **91** 11482–11487. https://doi.org/10.1021/acs.analchem.9b03340

RYU, S. Y., WENDT, G. A., ERNST, R. K. and GOODLETT, D. R. (2020). MGMS2: Membrane glycolipid mass spectrum simulator for polymicrobial samples. *Rapid Commun. Mass Spectrom.* **34** e8824. https://doi.org/10.1002/rcm.8824

VLEK, A. L., BONTEN, M. J. and BOEL, C. E. (2012). Direct matrix-assisted laser desorption ionization time-of-flight mass spectrometry improves appropriateness of antibiotic treatment of bacteremia. *PLoS ONE* **7** e32589.

YANG, Y., LIN, Y. and QIAO, L. (2018). Direct MALDI-TOF MS identification of bacterial mixtures. *Anal. Chem.* **90** 10400–10408.

YANG, Y., LIN, Y., CHEN, Z., GONG, T., YANG, P., GIRAULT, H., LIU, B. and QIAO, L. (2017). Bacterial whole cell typing by mass spectra pattern matching with bootstrapping assessment. *Anal. Chem.* **89** 12556–12561.

ZHANG, Z. (2004). Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76** 3908–3922.

The Annals of Applied Statistics
2022, Vol. 16, No. 1, 625
https://doi.org/10.1214/21-AOAS1573
Main article: https://doi.org/10.1214/20-AOAS1354

CORRECTION TO: A BAYESIAN MODEL OF MICROBIOME DATA FOR SIMULTANEOUS IDENTIFICATION OF COVARIATE ASSOCIATIONS AND PREDICTION OF PHENOTYPIC OUTCOMES

BY MATTHEW D. KOSLOVSKY[1,a], KRISTI L. HOFFMAN[2,c], CARRIE R. DANIEL[3,d] AND MARINA VANNUCCI[1,b]

[1]*Department of Statistics, Rice University,* [a]*mkoslovsky@rice.edu,* [b]*marina@rice.edu*
[2]*Alkek Center for Metagenomics & Microbiome Research, Baylor College of Medicine,* [c]*Kristi.Hoffman@bcm.edu*
[3]*Department of Epidemiology, The University of Texas MD Anderson Cancer Center,* [d]*CDaniel@mdanderson.org*

REFERENCES

KOSLOVSKY, M. D., HOFFMAN, K. L., DANIEL, C. R. and VANNUCCI, M. (2020). A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Ann. Appl. Stat.* **14** 1471–1492. MR4152142 https://doi.org/10.1214/20-AOAS1354