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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

continued

Vol. 19, No. 1—March 2025

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*

Vol. 19, No. 1—March 2025

THE ANNALS OF APPLIED STATISTICS Vol. 19, No. 1, pp. 1–845 March 2025

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.

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 19, Number 1, March 2025. 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
2025, Vol. 19, No. 1, 1–2
https://doi.org/10.1214/25-AOAS191ED

EDITORIAL: SHAPING THE FUTURE: A VISION FOR THE *Annals of Applied Statistics* IN A TRANSFORMATIVE ERA

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 3–27
https://doi.org/10.1214/24-AOAS1895

PERIODOGRAM REGRESSION: A TWO-STAGE MIXED EFFECTS APPROACH FOR MODELLING MULTIPLE INTEGER-VALUED TIME SERIES OF TROPICAL CYCLONE FREQUENCY

BY LYUYUAN ZHANG[1,a], GUOQI QIAN[1,b] AND SOURAV DAS[2,c]

[1]*School of Mathematics and Statistics, University of Melbourne,* [a]*lyuyuanz@student.unimelb.edu.au,* [b]*qguoqi@unimelb.edu.au*

[2]*College of Science and Engineering, James Cook University,* [c]*sourav.das@jcu.edu.au*

Tropical cyclones (TC) are significant indicators of evolving climate dynamics. Two primary responses of interest are the cyclone frequency and intensity. In this paper we propose a novel integrated modelling framework for simultaneous modelling of TC frequency across several meteorological regions within Australasia. The key methodological insight is to model the second-order properties of multiple integer-valued time series in frequency domain, instead of parametric time domain models. We take a two-stage semiparametric approach, where large scale environmental variation is modelled using generalized linear models while the stochastic variation, including spatial heterogeneity, is estimated using spectral analysis of time series under a hierarchical generating model. Using longitudinal data analysis, we are able to jointly model periodicities in TC frequencies and their correlation with El Niño–Southern Oscillation (ENSO) cycles as well as the spatial variation between regions. We project the fitted model to obtain one-step-ahead forecasts under the principle of best linear unbiased estimation. This semiparametric approach allows us to obliterate the uniqueness matter of parametric integer-valued time series modelling. Additional methodological advantages include tests for spatial heterogeneity and temporal second-order stationarity. The data analysis corroborates previous findings on declining trend of tropical cyclone frequencies, in the short-term.

REFERENCES

ARNELL, N. W. and GOSLING, S. N. (2016). The impacts of climate change on river flood risk at the global scale. *Clim. Change* **134** 387–401.

BERAN, J. (1994). *Statistics for Long-Memory Processes*. *Monographs on Statistics and Applied Probability* **61**. CRC Press, New York. MR1304490

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

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

BOM. What is a tropical cyclone. http://www.bom.gov.au/cyclone/tropical-cyclone-knowledge-centre/understanding/tc-info/.

BRESLOW, N. E. and CLAYTON, D. G. (1996). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CAMARGO, S. J., BARNSTON, A. G., KLOTZBACH, P. J. and LANDSEA, C. W. (2007). Seasonal tropical cyclone forecasts. *WMO Bull.* **56** 297–309.

CHAND, S. S., WALSH, K. J., CAMARGO, S. J., KOSSIN, J. P., TORY, K. J., WEHNER, M. F., CHAN, J. C., KLOTZBACH, P. J., DOWDY, A. J. et al. (2022). Declining tropical cyclone frequency under global warming. *Nat. Clim. Change* **12** 655–661.

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1339. MR1731494 https://doi.org/10.2307/2669946

Key words and phrases. Tropical cyclone frequency, periodogram, longitudinal data analysis, mixed effects modelling, time series forecast.

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

CRESSIE, N. 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

DAS, S., SUBBA RAO, S. and YANG, J. (2021). Spectral methods for small sample time series: A complete periodogram approach. *J. Time Series Anal.* **42** 597–621. MR4325667 https://doi.org/10.1111/jtsa.12584

DAVIS, R. A., DUNSMUIR, W. T. M. and STREETT, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika* **90** 777–790. MR2024757 https://doi.org/10.1093/biomet/90.4.777

DAVIS, R. A., DUNSMUIR, W. T. M. and WANG, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika* **87** 491–505. MR1789805 https://doi.org/10.1093/biomet/87.3.491

DAVIS, R. A. and MIKOSCH, T. (1999). The maximum of the periodogram of a non-Gaussian sequence. *Ann. Probab.* **27** 522–536. MR1681157 https://doi.org/10.1214/aop/1022677270

DAVIS, R. A., WANG, Y. and DUNSMUIR, W. T. M. (1999). Modeling time series of count data. In *Asymptotics, Nonparametrics, and Time Series. Statist. Textbooks Monogr.* **158** 63–113. Dekker, New York. MR1724696

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

DOWDY, A. J., QI, L., JONES, D., RAMSAY, H., FAWCETT, R. and KULESHOV, Y. (2012). Tropical cyclone climatology of the South Pacific Ocean and its relationship to El Niño-Southern Oscillation. *J. Climate* **25** 6108–6122.

FOKIANOS, K. and KEDEM, B. (2003). Regression theory for categorical time series. *Statist. Sci.* **18** 357–376. MR2061915 https://doi.org/10.1214/ss/1076102425

FOKIANOS, K. and TJØSTHEIM, D. (2011). Log-linear Poisson autoregression. *J. Multivariate Anal.* **102** 563–578. MR2755016 https://doi.org/10.1016/j.jmva.2010.11.002

FREEDMAN, D. and LANE, D. (1981). The empirical distribution of the Fourier coefficients of a sequence of independent, identically distributed long-tailed random variables. *Z. Wahrsch. Verw. Gebiete* **58** 21–39. MR0635269 https://doi.org/10.1007/BF00536193

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

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475 https://doi.org/10.1198/016214502760047113

GUILLAUMIN, A. P., SYKULSKI, A. M., OLHEDE, S. C., EARLY, J. J. and LILLY, J. M. (2017). Analysis of non-stationary modulated time series with applications to oceanographic surface flow measurements. *J. Time Series Anal.* **38** 668–710. MR3689441 https://doi.org/10.1111/jtsa.12244

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

HYNDMAN, R. J. and KHANDAKAR, Y. (2008). Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **26** 1–22.

JANAS, D. and VON SACHS, R. (1995). Consistency for non-linear functions of the periodogram of tapered data. *J. Time Series Anal.* **16** 585–606. MR1366801 https://doi.org/10.1111/j.1467-9892.1995.tb00257.x

KULESHOV, Y., QI, L., FAWCETT, R. and JONES, D. (2009). Improving preparedness to natural hazards: Tropical cyclone seasonal prediction for the Southern Hemisphere. In *Advances in Geosciences* **12** Ocean Science (OS) 127–143. World Scientific, Singapore.

KULESHOV, Y., WANG, Y., APAJEE, J., FAWCETT, R. and JONES, D. (2014). Prospects for improving the operational seasonal prediction of tropical cyclone activity in the Southern Hemisphere. In *Monitoring and Prediction of Tropical Cyclones in the Indian Ocean and Climate Change* 123–136. Springer, Berlin.

LIBOSCHIK, T., FOKIANOS, K. and FRIED, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *J. Stat. Softw.* **82** 1–51. https://doi.org/10.18637/jss.v082.i05

LÜDECKE, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *J. Open Sour. Softw.* **3** 772.

MAGEE, A. D. and KIEM, A. S. (2020). Using indicators of ENSO, IOD, and SAM to improve lead time and accuracy of tropical cyclone outlooks for Australia. *J. Appl. Meteorol. Climatol.* **59** 1901–1917.

MATÉRN, B. (1986). *Spatial Variation*, 2nd ed. *Lecture Notes in Statistics* **36**. Springer, Berlin. MR0867886 https://doi.org/10.1007/978-1-4615-7892-5

MCELROY, T. S. (2022). Frequency domain calculation of seasonal VARMA autocovariances. *J. Comput. Graph. Statist.* **31** 301–303. MR4387229 https://doi.org/10.1080/10618600.2021.1981343

MCPHADEN, M. J., ZEBIAK, S. E. and GLANTZ, M. H. (2006). ENSO as an integrating concept in Earth science. *Science* **314** 1740–1745. https://doi.org/10.1126/science.1132588

NICHOLLS, N. (1984). The Southern Oscillation, sea-surface-temperature, and interannual fluctuations in Australian tropical cyclone activity. *Int. J. Climatol.* **4** 661–670.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. and R CORE TEAM (2012). *Nlme*: *Linear and Nonlinear Mixed Effects Models R Foundation for Statistical Computing*, Vienna, Austria. R package version 3.1-104.

PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series*, *Vol*. 1. *Probability and Mathematical Statistics*. Academic Press, London. MR0628735

QIAN, G., CHEN, L. and KULESHOV, Y. (2022). Improving methodology for tropical cyclone seasonal forecasting in the Australian and the South Pacific Ocean regions by selecting and averaging models via Metropolis–Gibbs sampling. *Remote Sens*. **14** 5872.

R CORE TEAM (2017). *R*: *A Language and Environment for Statistical Computing R Foundation for Statistical Computing*, Vienna, Austria.

RAHIM, K. and BURR, W. S. (2017). multitaper: Spectral analysis tools using the multitaper method R package version 1.0-14. For new features, see the 'Changelog' file (in the package source).

RIEDEL, K. S. and SIDORENKO, A. (1995). Minimum bias multiple taper spectral estimation. *IEEE Trans*. *Signal Process*. **43** 188–195.

SARKAR, D. (2008). *Lattice*: *Multivariate Data Visualization with R*. Springer, Berlin.

SCHUSTER, A. (1906). The periodogram and its optical analogy. *Proc*. *R*. *Soc*. *Lond*. *Ser. A*, *Contain*. *Pap*. *Math*. *Phys*. *Character* **77** 136–140.

SENEVIRATNE, S., NICHOLLS, N., EASTERLING, D., GOODESS, C., KANAE, S., KOSSIN, J., LUO, Y., MARENGO, J., MCINNES, K. et al. (2012). Changes in climate extremes and their impacts on the natural physical environment.

SHUMWAY, R. H. and STOFFER, D. S. (2011). *Time Series Analysis and Its Applications*: *With R Examples*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2721825 https://doi.org/10.1007/978-1-4419-7865-3

THOMSON, D. J. (1982). Spectrum estimation and harmonic analysis. *Proc*. *IEEE* **70** 1055–1096.

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

VELASCO, C. (2000). Non-Gaussian log-periodogram regression. *Econometric Theory* **16** 44–79. MR1749019 https://doi.org/10.1017/S0266466600161031

WALDEN, A. T. (2000). A unified view of multitaper multivariate spectral estimation. *Biometrika* **87** 767–788. MR1813974 https://doi.org/10.1093/biomet/87.4.767

WEBSTER, P. J., HOLLAND, G. J., CURRY, J. A. and CHANG, H.-R. (2005). Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science* **309** 1844–1846.

WEISS, C. H. (2018). *An Introduction to Discrete-Valued Time Series*. Wiley, New York.

WEST, M. (1997). Time series decomposition. *Biometrika* **84** 489–494.

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449. MR0067450 https://doi.org/10.1093/biomet/41.3-4.434

WICKHAM, H. (2016). *ggplot2*: *Elegant Graphics for Data Analysis*. Springer, Berlin.

WIJNANDS, J., QIAN, G., SHELTON, K., FAWCETT, R., CHAN, J. and KULESHOV, Y. (2015). Seasonal forecasting of tropical cyclone activity in the Australian and the South Pacific Ocean regions. *Math*. *Clim*. *Weather Forecast*. **1**.

WIJNANDS, J., SHELTON, K. and KULESHOV, Y. (2014). Improving the operational methodology of tropical cyclone seasonal prediction in the Australian and the South Pacific Ocean regions. *Adv. Meteorol*. **2014**.

WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2206355

WOOD, S. N. (2013). A simple test for random effects in regression models. *Biometrika* **100** 1005–1010. MR3142347 https://doi.org/10.1093/biomet/ast038

ZHANG, L., QIAN, G. and DAS, S. (2025). Supplement to "Periodogram regression: A two-stage mixed effects approach for modelling multiple integer-valued time series of tropical cyclone frequency." https://doi.org/10.1214/24-AOAS1895SUPP

ZHAO-GUO, C. and HANNAN, E. J. (1980). The distribution of periodogram ordinates. *J. Time Series Anal*. **1** 73–82. MR0605575 https://doi.org/10.1111/j.1467-9892.1980.tb00301.x

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 28–55
https://doi.org/10.1214/24-AOAS1913

HETEROGENEOUS TREATMENT AND SPILLOVER EFFECTS UNDER CLUSTERED NETWORK INTERFERENCE

BY FALCO J. BARGAGLI-STOFFI[1,a], COSTANZA TORTÚ[2,b] AND LAURA FORASTIERE[3,c]

[1]*Department of Biostatistics, University of California, Los Angeles,* [a]*falco@g.ucla.edu*
[2]*Management and Healthcare Laboratory, Sant'Anna School for Advanced Studies,* [b]*costanza.tortu@santannapisa.it*
[3]*Department of Biostatistics, Yale University,* [c]*laura.forastiere@yale.edu*

The bulk of causal inference studies rule out the presence of interference between units. However, in many real-world scenarios, units are interconnected by social, physical, or virtual ties, and the effect of the treatment can spill from one unit to other connected individuals in the network. In this paper, we develop a machine learning method that uses tree-based algorithms and a Horvitz–Thompson estimator to assess the heterogeneity of treatment and spillover effects with respect to individual, neighborhood, and network characteristics in the context of clustered networks and interference within clusters. The proposed network causal tree (NCT) algorithm has several advantages. First, it allows the investigation of the heterogeneity of the treatment effect, avoiding potential bias due to the presence of interference. Second, understanding the heterogeneity of both treatment and spillover effects can guide policymakers in scaling up interventions, designing targeting strategies, and increasing cost-effectiveness. We investigate the performance of our NCT method using a Monte Carlo simulation study and illustrate its application to assess the heterogeneous effects of information sessions on the uptake of a new weather insurance policy in rural China.

REFERENCES

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

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BARGAGLI-STOFFI, F. J., CADEI, R., LEE, K. and DOMINICI, F. (2020). Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. ArXiv preprint. Available at arXiv:2009.09036.

BARGAGLI-STOFFI, F. J., DE WITTE, K. and GNECCO, G. (2022). Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach. *Ann. Appl. Stat.* **16** 1986–2009. MR4455908 https://doi.org/10.1214/21-aoas1579

BARGAGLI-STOFFI, F. J. and GNECCO, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. *Int. J. Data Sci. Anal.* **9** 315–337.

BARGAGLI-STOFFI, F. J., TORTÚ, C. and FORASTIERE, L. (2025). Supplement to "Heterogeneous treatment and spillover effects under clustered network interference." https://doi.org/10.1214/24-AOAS1913SUPP

BARNETT, B. J. and MAHUL, O. (2007). Weather index insurance for agriculture and rural areas in lower-income countries. *Amer. J. Agric. Econ.* **89** 1241–1247.

CAI, J., DE JANVRY, A. and SADOULET, E. (2015). Social networks and the decision to insure. *Am. Econ. J. Appl. Econ.* **7** 81–108.

CAI, J. and SONG, C. (2017). Do disaster experience and knowledge affect insurance take-up decisions? *J. Dev. Econ.* **124** 83–94.

Key words and phrases. Causal inference, interpretable machine learning, interference, social networks, heterogeneous effects.

CENTRE FOR RESEARCH ON THE EPIDEMIOLOGY OF DISASTERS (CRED) (2022). Disaster Year in Review (2021). CRED Crunch 60, Center for the Research on the Epidemiology of Disasters, Brussels, Belgium.

CHAE, I., STEPHEN, A. T., BART, Y. and YAO, D. (2017). Spillover effects in seeded word-of-mouth marketing campaigns. *Mark. Sci.* **36** 89–104.

CHAKRABORTY, B. and MURPHY, S. (2014). Dynamic treatment regimes. *Annu. Rev. Stat. Appl.* **1** 447–464.

CHEN, L. H. Y. and SHAO, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028. MR2073183 https://doi.org/10.1214/009117904000000450

COLE, S., GINÉ, X., TOBACMAN, J., TOWNSEND, R., TOPALOVA, P. and VICKERY, J. (2013). Barriers to household risk management: Evidence from India. *Am. Econ. J. Appl. Econ.* **5** 104–135. https://doi.org/10.1257/app.5.1.104

COX, D. R. (1958). *Planning of Experiments. A Wiley Publication in Applied Statistics*. Wiley, New York. MR0095561

DERCON, S., HILL, R. V., CLARKE, D., OUTES-LEON, I. and TAFFESSE, A. S. (2014). Offering rainfall insurance to informal insurance groups: Evidence from a field experiment in Ethiopia. *J. Dev. Econ.* **106** 132–143.

EBI, K. L., VANOS, J., BALDWIN, J. W., BELL, J. E., HONDULA, D. M., ERRETT, N. A., HAYES, K., REID, C. E., SAHA, S. et al. (2021). Extreme weather and climate change: Population health and health system implications. *Annu. Rev. Public Health* **42** 293.

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. MR0120167 https://doi.org/10.5486/pmd.1959.6.3-4.12

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

FORTUNATO, S. (2010). Community detection in graphs. *Phys. Rep.* **486** 75–174. MR2580414 https://doi.org/10.1016/j.physrep.2009.11.002

FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J. et al. (1984). *Classification and Regression Trees*. Wadsworth & Brooks, Belmont, CA.

GAURAV, S., COLE, S. and TOBACMAN, J. (2011). The randomized evaluation of financial literacy on rainfall insurance take-up in Gujarat. *ILO Microins. Innov. Facil. Res. Paper* **1**.

GINÉ, X., TOWNSEND, R. and VICKERY, J. (2008). Patterns of rainfall insurance participation in rural India. *World Bank Econ. Rev.* **22** 539–566.

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

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

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

HUQ, N., HUGÉ, J., BOON, E. and GAIN, A. K. (2015). Climate change impacts in agricultural communities in rural areas of coastal Bangladesh: A tale of many stories. *Sustainability* **7** 8437–8460.

KEMPE, D., KLEINBERG, J. and TARDOS, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137–146.

KENT, D. M., PAULUS, J. K., VAN KLAVEREN, D., D'AGOSTINO, R., GOODMAN, S., HAYWARD, R., IOANNIDIS, J. P., PATRICK-LAKE, B., MORTON, S. et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann. Intern. Med.* **172** 35–45.

KIM, D., HWONG, A., STAFFORD, D., HUGHES, D., O'MALLEY, J., FOWLER, J. and CHRISTAKIS, N. (2015). Social network targeting to maximise population behaviour change: A cluster randomised controlled trial. *Lancet* **386** 145–153.

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

LAL, P., ALAVALAPATI, J. R. and MERCER, E. D. (2011). Socio-economic impacts of climate change on rural United States. *Mitigation and Adaptation Strategies for Global Change* **16** 819–844.

LIU, B., LI, M., GUO, Y. and SHAN, K. (2010). Analysis of the demand for weather index agricultural insurance on household level in Anhui, China. *Agric. Agric. Sci. Proc.* **1** 179–186.

MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060 https://doi.org/10.1111/j.1368-423X.2012.00368.x

MONTES, F., JARAMILLO, A. M., MEISEL, J. D., DIAZ-GUILERA, A., VALDIVIA, J. A., SARMIENTO, O. L. and ZARAMA, R. (2020). Benchmarking seeding strategies for spreading processes in social networks: An interplay between influencers, topologies and sizes. *Sci. Rep.* **10** 3666.

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

PARSHAKOV, P., NAIDENOVA, I. and BARAJAS, A. (2020). Spillover effect in promotion: Evidence from video game publishers and eSports tournaments. *J. Bus. Res.* **118** 262–270.

ROGERS, S. and XUE, T. (2015). Resettlement and climate change vulnerability: Evidence from rural China. *Glob. Environ. Change* **35** 62–69.

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

RUBIN, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments [MR2655714. *J. Amer. Statist. Assoc.* **103** 1350–1353. MR2655717 https://doi.org/10.1198/016214508000001011

SÄVJE, F. (2024). Causal inference with misspecified exposure mappings: Separating definitions and assumptions. *Biometrika* **111** 1–15. MR4704553 https://doi.org/10.1093/biomet/asad019

SIBIKO, K. W. and QAIM, M. (2020). Weather index insurance, agricultural input use, and crop productivity in Kenya. *Food Secur.* **12** 151–167.

VALENTE, T. W. (2012). Network interventions. *Science* **337** 49–53. https://doi.org/10.1126/science.1217330

VANDERWEELE, T. J. and CHRISTAKIS, N. A. (2019). Network multipliers and public health. *Int. J. Epidemiol.* **48** 1032–1037.

VIVIANO, D. (2024). Policy targeting under network interference. *Rev. Econ. Stud.*

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

ZHAO, W. (2020). Extreme weather and climate events in China under changing climate. *Nat. Sci. Rev.* **7** 938–943.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 56–74
https://doi.org/10.1214/24-AOAS1914

LEVERAGING CELLPHONE-DERIVED MOBILITY NETWORKS TO ASSESS COVID-19 TRAVEL RISK

BY JUSTIN J. SLATER[1,a], PATRICK E. BROWN[2,b], JEFFREY S. ROSENTHAL[2,c] AND JORGE MATEU[3,d]

[1]*Department of Mathematics and Statistics, University of Guelph,* [a]*jslate04@uoguelph.ca*

[2]*Department of Statistical Sciences, University of Toronto,* [b]*patrick.brown@utoronto.ca,* [c]*jeff@math.toronto.edu*

[3]*Department of Mathematics, University Jaume I of Castellon,* [d]*mateu@mat.uji.es*

Since the beginning of the Covid-19 pandemic, public health authorities across the globe have implemented policies, such as lockdowns, in an attempt to reduce population mobility and, consequently, person-to-person contacts. It is well known that lockdowns reduce mobility, but to what extent does this reduction in mobility lead to lower infection rates? In this paper we extend the endemic-epidemic modeling framework in a principled manner, incorporating temporally changing mobility network data and quantifying the risk associated with travelling throughout the first year of the pandemic in two Spanish communities.

REFERENCES

ALENE, M., YISMAW, L., ASSEMIE, M. A., KETEMA, D. B., GIETANEH, W. and BIRHAN, T. Y. (2021). Serial interval and incubation period of COVID-19: A systematic review and meta-analysis. *BMC Infect. Dis.* **21** 1–9.

BAUER, C. and WAKEFIELD, J. (2018). Stratified space-time infectious disease modelling, with an application to hand, foot and mouth disease in China. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1379–1398. MR3873712 https://doi.org/10.1111/rssc.12284

BRACHER, J. and HELD, L. (2022). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *Int. J. Forecast.* **38** 1221–1233.

BRACHER, J. and HELD, L. (2021). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts. *Biometrics* **77** 1202–1214. MR4357831 https://doi.org/10.1111/biom.13371

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

CELANI, A. and GIUDICI, P. (2022). Endemic-epidemic models to understand COVID-19 spatio-temporal evolution. *Spat. Stat.* **49** 100528. MR4439304 https://doi.org/10.1016/j.spasta.2021.100528

DIEKMANN, O., HEESTERBEEK, H. and BRITTON, T. (2013). *Mathematical Tools for Understanding Infectious Disease Dynamics*. *Princeton Series in Theoretical and Computational Biology*. Princeton Univ. Press, Princeton, NJ. MR3015083

DIEKMANN, O., HEESTERBEEK, J. A. P. and METZ, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28** 365–382. MR1057044 https://doi.org/10.1007/BF00178324

DOUWES-SCHULTZ, D., SUN, S., SCHMIDT, A. M. and MOODIE, E. E. M. (2022). Extended Bayesian endemic-epidemic models to incorporate mobility data into COVID-19 forecasting. *Canad. J. Statist.* **50** 713–733. MR4482638 https://doi.org/10.1002/cjs.11723

FRITZ, C. and KAUERMANN, G. (2022). On the interplay of regional mobility, social connectedness and the spread of COVID-19 in Germany. *J. Roy. Statist. Soc. Ser. A* **185** 400–424. MR4384311 https://doi.org/10.1111/rssa.12753

GARCÍA-GARCÍA, D., VIGO, M. I., FONFRÍA, E. S., HERRADOR, Z., NAVARRO, M. and BORDEHORE, C. (2021). Retrospective methodology to estimate daily infections from deaths (REMEDID) in COVID-19: The Spain case study. *Sci. Rep.* **11** 1–15.

GEILHUFE, M., HELD, L., SKRØVSETH, S. O., SIMONSEN, G. S. and GODTLIEBSEN, F. (2014). Power law approximations of movement network data for modeling infectious disease spread. *Biom. J.* **56** 363–382. MR3256495 https://doi.org/10.1002/bimj.201200262

Key words and phrases. Cellphone mobility data, Covid-19, infectious diseases, endemic-epidemic models, Poisson network autoregressions, spatiotemporal statistics.

GENERAL DIRECTORATE OF INFORMATION SYSTEMS, QUALITY AND PHARMACEUTICAL PROVISION (2022). Open data of Castile and Leon. https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html. Accessed 2022-03-10.

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

GORDIS, L. (2013). *Epidemiology*, 5th ed. Elsevier Health Sciences, Philadelphia, PA.

GOTTUMUKKALA, R., KATRAGADDA, S., BHUPATIRAJU, R. T., KAMAL, A. M., RAGHAVAN, V., CHU, H., KOLLURU, R. and ASHKAR, Z. (2021). Exploring the relationship between mobility and COVID-19 infection rates for the second peak in the United States using phase-wise association. *BMC Public Health* **21** 1–14.

GRIMÉE, M., BEKKER-NIELSEN DUNBAR, M., HOFMANN, F. and HELD, L. (2022). Modelling the effect of a border closure between Switzerland and Italy on the spatiotemporal spread of COVID-19 in Switzerland. *Spat. Stat.* **49** 100552. MR4439317 https://doi.org/10.1016/j.spasta.2021.100552

HALE, T., ANGRIST, N., GOLDSZMIDT, R., KIRA, B., PETHERICK, A., PHILLIPS, T., WEBSTER, S., CAMERON-BLAKE, E., HALLAS, L. et al. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5** 529–538.

HALLORAN, M. E., LONGINI, I. M. JR. and STRUCHINER, C. J. (2010). *Design and Analysis of Vaccine Studies*. *Statistics for Biology and Health*. Springer, New York. MR2572061 https://doi.org/10.1007/978-0-387-68636-3

HE, X., LAU, E. H., WU, P., DENG, X., WANG, J., HAO, X., LAU, Y. C., WONG, J. Y., GUAN, Y. et al. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26** 672–675.

HELD, L., HOFMANN, M., HÖHLE, M. and SCHMID, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics* **7** 422–437. https://doi.org/10.1093/biostatistics/kxj016

HELD, L., HÖHLE, M. and HOFMANN, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Stat. Model.* **5** 187–199. MR2210732 https://doi.org/10.1191/1471082X05st098oa

HELD, L. and PAUL, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biom. J.* **54** 824–843. MR2993630 https://doi.org/10.1002/bimj.201200037

KNORR-HELD, L. and RICHARDSON, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *J. R. Stat. Soc., Ser. C* **52** 169–183. MR1973739 https://doi.org/10.1111/1467-9876.00396

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

MEYER, S. and HELD, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics* **18** 338–351. MR3824756 https://doi.org/10.1093/biostatistics/kxw051

MINISTERIO DE SANIDAD, GOBIERNO DE ESPAÑA (2022a). Estrategia de Vacunacion COVID-19 EN España. https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/vacunaCovid19.htm. Accessed 2022-05-17.

MINISTERIO DE SANIDAD, GOBIERNO DE ESPAÑA (2022b). Datos abiertos de pruebas realizadas. https://www.sanidad.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/pruebasRealizadas.htm. Accessed 2022-05-17.

MINISTERIO DE TRANSPORTES MOVILIDAD Y AGENDS URBANA, GOBIERNO DE ESPAÑA (2022). Evolución de la movilidad diara. www.mitma.gob.es/ministerio/covid-19/evolucion-movilidad-big-data. Accessed 2022-10-26.

EPIDEMIOLOGICAL SURVEILLANCE NETWORK OF MADRID (2022). COVID 19-TIA by municipalities and districts of Madrid. https://datos.gob.es/en/catalogo/a13002908-covid-19-tia-por-municipios-y-distritos-de-madrid1. Accessed 2022-05-17.

PAUL, M. and HELD, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat. Med.* **30** 1118–1136. MR2767846 https://doi.org/10.1002/sim.4177

PAUL, M., HELD, L. and TOSCHKE, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Stat. Med.* **27** 6250–6267. MR2522320 https://doi.org/10.1002/sim.3440

PONCE-DE LEON, M., DEL VALLE, J., FERNANDEZ, J. M., BERNARDO, M., CIRILLO, D., SANCHEZ-VALLE, J., SMITH, M., CAPELLA-GUTIERREZ, S., GULLÓN, T. et al. (2021). COVID-19 flow-maps an open geographic information system on COVID-19 and human mobility for Spain. *Sci. Data* **8** 1–16.

SCHRÖDLE, B., HELD, L. and RUE, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68** 736–744. MR3055178 https://doi.org/10.1111/j.1541-0420.2011.01717.x

SLATER, J. J., BROWN, P. E. and ROSENTHAL, J. S. (2021). Forecasting subnational COVID-19 mortality using a day-of-the-week adjusted Bayesian hierarchical model. *Stat* **10** e328. MR4235597 https://doi.org/10.1002/sta4.328

SLATER, J. J., BROWN, P. E., ROSENTHAL, J. S. and MATEU, J. (2022). Capturing spatial dependence of COVID-19 case counts with cellphone mobility data. *Spat*. *Stat*. **49** 100540. MR4439309 https://doi.org/10.1016/j.spasta.2021.100540

SLATER, J. J., BROWN, P. E., ROSENTHAL, J. S. and MATEU, J. (2025). Supplement to "Leveraging cellphone-derived mobility networks to assess Covid-19 travel risk." https://doi.org/10.1214/24-AOAS1914SUPP

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

VEHTARI, A., GABRY, J., MAGNUSSON, M., YAO, Y., BÜRKNER, P.-C., PAANANEN, T. and GELMAN, A. (2024). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.7.0.

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal*. **16** 667–718. MR4298989 https://doi.org/10.1214/20-ba1221

WAKEFIELD, J., DONG, T. Q. and MININ, V. N. (2019). Spatio-temporal analysis of surveillance data. In *Handbook of Infectious Disease Data Analysis* 455–475. CRC Press/CRC, Boca Raton.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 75–93
https://doi.org/10.1214/24-AOAS1918

RISK SET MATCHED DIFFERENCE-IN-DIFFERENCES FOR THE ANALYSIS OF EFFECT MODIFICATION IN AN OBSERVATIONAL STUDY ON THE IMPACT OF GUN VIOLENCE ON HEALTH OUTCOMES

BY ERIC R. COHN[1,a], ZIRUI SONG[2,b] AND JOSÉ R. ZUBIZARRETA[3,c]

[1]*Department of Biostatistics, Harvard School of Public Health,* [a]*ericcohn@g.harvard.edu*

[2]*Department of Health Care Policy, Harvard University,* [b]*song@hcp.med.harvard.edu*

[3]*Departments of Health Care Policy, Biostatistics, and Statistics, Harvard University,* [c]*zubizarreta@hcp.med.harvard.edu*

Gun violence is a major source of injury and death in the United States. However, relatively little is known about the effects of firearm injuries on survivors and their family members and how these effects vary. To study these questions and, more generally, to address a gap in the methodological causal inference literature, we present a framework for the study of effect modification or heterogeneous treatment effects in difference-in-differences designs. We implement a new matching technique, combining profile matching and risk set matching to: (i) preserve the time alignment of covariates, exposure, and outcomes, avoiding pitfalls of other common approaches for difference-in-differences, and (ii) explicitly control biases due to imbalances in observed covariates in subgroups discovered from the data. Our case study shows significant and persistent effects of nonfatal firearm injuries on several health outcomes for those injured and on the mental health of their family members. Sensitivity analyses reveal that these results are moderately robust to unmeasured confounding bias. Finally, while the effects for those injured vary largely by the severity of the injury and its documented intent, for families, effects are strongest for those whose relative's injury is documented as resulting from an assault, self-harm, or law enforcement intervention.

REFERENCES

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

ATHEY, S. and IMBENS, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74** 431–497. MR2207397 https://doi.org/10.1111/j.1468-0262.2006.00668.x

ATHEY, S. and IMBENS, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *J. Econometrics* **226** 62–79. MR4348786 https://doi.org/10.1016/j.jeconom.2020.10.012

BAKER, A. C., LARCKER, D. F. and WANG, C. C. Y. (2022). How much should we trust staggered difference-in-differences estimates? *J. Financ. Econ.* **144** 370–395.

BASU, P. and SMALL, D. S. (2020). Constructing a more closely matched control group in a difference-in-differences analysis: Its effect on history interacting with group bias. *Obs. Stud.* **6** 103–130.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2022). Synthetic controls with staggered adoption. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 351–381. MR4412990

BLACKWELL, M. and STREZHNEV, A. (2022). Telescope matching for reducing model dependence in the estimation of the effects of time-varying treatments: An application to negative advertising. *J. Roy. Statist. Soc. Ser. A* **185** 377–399. MR4384310 https://doi.org/10.1111/rssa.12759

BORUSYAK, K., JARAVEL, X. and SPIESS, J. (2021). Revisiting event study designs: Robust and efficient estimation. arXiv preprint. Available at arXiv:2205.08644. https://doi.org/10.48550/ARXIV.2108.12419

CALLAWAY, B. and SANT'ANNA, P. H. C. (2021). Difference-in-differences with multiple time periods. *J. Econometrics* **225** 200–230. MR4328640 https://doi.org/10.1016/j.jeconom.2020.12.001

CARD, D. and KRUEGER, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *Amer. Econ. Rev.* **84** 772–793.

Key words and phrases. Causal inference, difference-in-differences, effect modification, multivariate matching, observational studies.

CHARTERS, T. J., STRUMPF, E. C. and SEWITCH, M. J. (2013). Effectiveness of an organized colorectal cancer screening program on increasing adherence in asymptomatic average-risk Canadians. *BMC Health Serv. Res.* **13** 449–449.

COHN, E. R., SONG, Z. and ZUBIZARRETA, J. R. (2025). Supplement to "Risk Set Matched Difference-in-Differences for the Analysis of Effect Modification in an Observational Study on the Impact of Gun Violence on Health Outcomes." https://doi.org/10.1214/24-AOAS1918SUPP

COHN, E. R. and ZUBIZARRETA, J. R. (2022). Profile matching for the generalization and personalization of causal inferences. *Epidemiology* **33** 678–688. https://doi.org/10.1097/EDE.0000000000001517

DAW, J. R. and HATFIELD, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Serv. Res.* **53** 4138–4156.

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

DING, P., FELLER, A. and MIRATRIX, L. (2019). Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* **114** 304–317. MR3941256 https://doi.org/10.1080/01621459.2017.1407322

DING, P. and LI, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Polit. Anal.* **27** 605–615.

FITZPATRICK, V., CASTRO, M., JACOBS, J., SEBRO, N., GULMATICO, J., SHIELDS, M. and HOMAN, S. M. (2019). Nonfatal firearm violence trends on the westside of Chicago between 2005 and 2016. *J. Commun. Health* **44** 866–873. https://doi.org/10.1007/s10900-018-00603-8

FOGARTY, C. B. (2023). Testing weak nulls in matched observational studies. *Biometrics* **79** 2196–2207. MR4643987 https://doi.org/10.1111/biom.13741

FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.* **111** 1820–1830. MR3601738 https://doi.org/10.1080/01621459.2015.1120675

FOWLER, K. A., DAHLBERG, L. L., HAILEYESUS, T. and ANNEST, J. L. (2015). Firearm injuries in the United States. *Prev. Med.* **79** 5–14. https://doi.org/10.1016/j.ypmed.2015.06.002

FREDRIKSSON, A. and OLIVEIRA, G. M. D. (2019). Impact evaluation using difference-in-differences. *RAUSP Manag. J.* **54** 519–532.

GOODMAN-BACON, A. (2021). Difference-in-differences with variation in treatment timing. *J. Econometrics* **225** 254–277. MR4328642 https://doi.org/10.1016/j.jeconom.2021.03.014

HAVILAND, A., NAGIN, D. S. and ROSENBAUM, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychol. Methods* **12** 247–267.

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

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

IMAI, K., KIM, I. S. and WANG, E. H. (2021). Matching methods for causal inference with time-series cross-sectional data. *Amer. J. Polit. Sci.*

IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. MR2427345 https://doi.org/10.1111/j.1467-985X.2007.00527.x

KALESAN, B., ADHIKARLA, C., PRESSLEY, J. C., FAGAN, J. A., XUAN, Z., SIEGEL, M. B. and GALEA, S. (2017). The hidden epidemic of firearm injury: Increasing firearm injury rates during 2001–2013. *Amer. J. Epidemiol.* **185** 546–553. https://doi.org/10.1093/aje/kww147

KAUFMAN, E. J., WIEBE, D. J., XIONG, R. A., MORRISON, C. N., SEAMON, M. J. and DELGADO, M. K. (2021). Epidemiologic trends in fatal and nonfatal firearm injuries in the US, 2009–2017. *Arch. Intern. Med.* **181** 237–244.

KEELE, L., HASEGAWA, R. and SMALL, D. (2019). Bracketing bounds for differences-in-differences with an application to voter ID laws. MIT working paper.

KEELE, L. J., SMALL, D. S., HSU, J. Y. and FOGARTY, C. B. (2019). Patterns of effects and sensitivity analysis for differences-in-differences. arXiv preprint. Available at arXiv:1901.01869. https://doi.org/10.48550/ARXIV.1901.01869

LEE, K., BARGAGLI-STOFFI, F. J. and DOMINICI, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. arXiv preprint. Available at arXiv:2009.09036. https://doi.org/10.48550/ARXIV.2009.09036

LEE, K., SMALL, D. S. and DOMINICI, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *J. Amer. Statist. Assoc.* **116** 569–580. MR4270004 https://doi.org/10.1080/01621459.2020.1870476

LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134 https://doi.org/10.1111/biom.12884

LEIFHEIT, K. M., LINTON, S. L., RAIFMAN, J., SCHWARTZ, G. L., BENFER, E. A., ZIMMERMAN, F. J. and POLLACK, C. E. (2021). Expiring eviction moratoriums and COVID-19 incidence and mortality. *Amer. J. Epidemiol.* **190** 2563–2570.

LI, Y. P., PROPERT, K. J. and ROSENBAUM, P. R. (2001). Balanced risk set matching. *J. Amer. Statist. Assoc.* **96** 870–882. MR1946360 https://doi.org/10.1198/016214501753208573

LONG, S. K., YEMANE, A. and STOCKLEY, K. (2010). Disentangling the effects of health reform in Massachusetts: How important are the special provisions for young adults? *Amer. Econ. Rev.* **100** 297–302. https://doi.org/10.1257/aer.100.2.297

MAGEE, L. A., AALSMA, M. C., FORTENBERRY, J. D., GHARBI, S. and WIEHE, S. E. (2022). Mental health outcomes from direct and indirect exposure to firearm violence: A cohort study of nonfatal shooting survivors and family members. *J. Crim. Justice* **83**.

MARCUS, R., ERIC, P. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

MING, K. and ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56** 118–124. https://doi.org/10.1111/j.0006-341x.2000.00118.x

NEYMAN, J. S. DABROWSKA, D. M. SPEED, T. P. (1990). On the application of probability theory to agricultural experiments. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dąbrowska and T. P. Speed. MR1092986

RIGDON, J., BAIOCCHI, M. and BASU, S. (2018). Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials* **19** 382–382.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804 https://doi.org/10.1111/j.1541-0420.2006.00717.x

ROSENBAUM, P. R. (2018). Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Ann. Appl. Stat.* **12** 2312–2334. MR3875702 https://doi.org/10.1214/18-AOAS1153

ROSENBAUM, P. R. (2020). *Design of Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, Cham. MR4225301 https://doi.org/10.1007/978-3-030-46405-9

ROSENBAUM, P. R. and SILBER, J. H. (2009). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Amer. Statist. Assoc.* **104** 501–511. MR2751434 https://doi.org/10.1198/jasa.2009.0016

ROTH, J. and SANT'ANNA, P. H. C. (2023). When is parallel trends sensitive to functional form? *Econometrica* **91** 737–747. MR4566237

ROTH, J., SANT'ANNA, P. H. C., BILINSKI, A. and POE, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *J. Econometrics* **235** 2218–2244. MR4602950 https://doi.org/10.1016/j.jeconom.2023.03.008

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

SCHUMACHER, S., KIRZINGER, A., PRESIADO, M., VALDES, I. and BRODIE, M. (2023). Americans' experiences with gun-related violence, injuries, and deaths. Available at https://www.kff.org/other/poll-finding/americans-experiences-with-gun-related-violence-injuries-and-deaths/.

SNOW, J. (1856). On the mode of communication of cholera. *Edinb. Med. J.* **1** 668–670.

SOFER, T., RICHARDSON, D. B., COLICINO, E., SCHWARTZ, J. and TCHETGEN TCHETGEN, E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statist. Sci.* **31** 348–361. MR3552739 https://doi.org/10.1214/16-STS558

SONG, Z. (2022). The business case for reducing firearm injuries. *J. Amer. Med. Assoc.* **328** 1185–1186.

SONG, Z., ZUBIZARRETA, J. R., GIURIATO, M., PAULOS, E. and KOH, K. A. (2022). Changes in health care spending, use, and clinical outcomes after nonfatal firearm injuries among survivors and family members: A cohort study. *Ann. Intern. Med.* **175** 795–803. https://doi.org/10.7326/M21-2812

SOUTH, E. C., MACDONALD, J. M., TAM, V. W., RIDGEWAY, G. and BRANAS, C. C. (2022). Effect of abandoned housing interventions on gun violence, perceptions of safety, and substance use in black neighborhoods: A citywide cluster randomized trial. *JAMA Intern. Med.* https://doi.org/10.1001/jamainternmed.2022.5460

STUART, E. A., HUSKAMP, H. A., DUCKWORTH, K., SIMMONS, J., SONG, Z., CHERNEW, M. E. and BARRY, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Serv. Outcomes Res. Methodol.* **14** 166–182.

SUBBARAMAN, N. (2021). Gun violence is surging - researchers finally have the money to ask why. *Nature* **595** 486–488. https://doi.org/10.1038/d41586-021-01966-0

SUN, L. and ABRAHAM, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econometrics* **225** 175–199. MR4328639 https://doi.org/10.1016/j.jeconom.2020.09.006

WINTEMUTE, G. J. (2015). The epidemiology of firearm violence in the twenty-first century United States. *Annu. Rev. Public Health* **36** 5–19. https://doi.org/10.1146/annurev-publhealth-031914-122535

YANG, J., DAHABREH, I. J. and STEINGRIMSSON, J. A. (2022). Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics* **78** 624–635. MR4450581 https://doi.org/10.1111/biom.13432

YE, T., ERTEFAIE, A., FLORY, J., HENNESSY, S. and SMALL, D. S. (2023). Instrumented difference-in-differences. *Biometrics* **79** 569–581. MR4606300 https://doi.org/10.1111/biom.13783

YU, W., GUO, Y., SHI, L. and LI, S. (2020). The association between long-term exposure to low-level PM2.5 and mortality in the state of Queensland, Australia: A modelling study with the difference-in-differences approach. *PLoS Med.* **17** e1003141–e1003141.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 94–120
https://doi.org/10.1214/24-AOAS1929

PREDICTING CENSUS SURVEY RESPONSE RATES WITH PARSIMONIOUS ADDITIVE MODELS AND STRUCTURED INTERACTIONS

BY SHIBAL IBRAHIM[1,a], PETER RADCHENKO[2,b], EMANUEL BEN-DAVID[3,c] AND RAHUL MAZUMDER[4,d]

[1]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,* [a]*shibal@mit.edu*

[2]*University of Sydney Business School, University of Sydney,* [b]*peter.radchenko@sydney.edu.au*

[3]*Center for Statistical Research and Methodology, United States Census Bureau,* [c]*emanuel.ben.david@census.gov*

[4]*MIT Sloan School of Management, Massachusetts Institute of Technology,* [d]*rahulmaz@mit.edu*

In this paper we consider the problem of predicting survey response rates using a family of flexible and interpretable nonparametric models. The study is motivated by the U.S. Census Bureau's well-known ROAM application, which uses a linear regression model trained on the U.S. Census Planning Database data to identify hard-to-survey areas. A crowdsourcing competition (*Public Opin. Q.* **81** (2016) 144–156) organized more than 10 years ago revealed that machine learning methods, based on ensembles of regression trees, led to the best performance in predicting survey response rates; however, the corresponding models could not be adopted for the intended application due to their black-box nature. We consider nonparametric additive models with a small number of main and pairwise interaction effects using ℓ_0-based penalization. From a methodological viewpoint, we study our estimator's computational and statistical aspects and discuss variants incorporating strong hierarchical interactions. Our algorithms (open-sourced on GitHub) extend the computational frontiers of existing algorithms for sparse additive models to be able to handle datasets relevant to the application we consider. We discuss and interpret findings from our model on the U.S. Census Planning Database. In addition to being useful from an interpretability standpoint, our models lead to predictions comparable to popular black-box machine learning methods based on gradient boosting and feedforward neural networks—suggesting that it is possible to have models that have the best of both worlds, good model accuracy and interpretability.

REFERENCES

BACH, F., JENATTON, R., MAIRAL, J. and OBOZINSKI, G. (2012). Structured sparsity through convex optimization. *Statist. Sci.* **27** 450–468. MR3025128 https://doi.org/10.1214/12-STS394

BATES, N. and MULRY, M. H. (2011). Using a geographic segmentation to understand, predict, and plan for census and survey mail nonresponse. *J. Off. Stat.* **27** 601–618.

BERGSTRA, J., KOMER, B., ELIASMITH, C. et al. (2015). Hyperopt: A Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8** 014008.

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

BRUCE, A., ROBINSON, J. G. and SANDERS, M. V. (2001). Hard-to-count scores and broad demographic groups associated with patterns of response rates in Census 2000. In *Joint Statistical Meeting*.

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

CHEN, T. and GUESTRIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD* '16 785–794. ACM, New York, NY, USA.

Key words and phrases. Population surveys, low-response score, nonparametric additive models with interactions, ℓ_0 sparsity, hierarchical sparsity, large-scale optimization, integer programming.

ERDMAN, C. and BATES, N. J. (2014). The U.S. Census Bureau mail return rate challenge: Crowdsourcing to develop a hard-to-count score.

ERDMAN, C. and BATES, N. (2016). The Low Response Score (LRS). *Public Opin*. *Q*. **81** 144–156.

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

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann*. *Statist*. **19** 1–67.

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

HASTIE, T. and TIBSHIRANI, R. (1987). Generalized additive models: Some applications. *J*. *Amer*. *Statist*. *Assoc*. **82** 371–386.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*. *Springer Series in Statistics*. Springer, New York. MR1851606 https://doi.org/10.1007/978-0-387-21606-5

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HAZIMEH, H. and MAZUMDER, R. (2020a). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Oper*. *Res*. **68** 1517–1537. MR4166310

HAZIMEH, H. and MAZUMDER, R. (2020b). Learning hierarchical interactions at scale: A convex optimization approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 1833–1843. PMLR.

HAZIMEH, H., MAZUMDER, R. and RADCHENKO, P. (2023). Grouped variable selection with discrete optimization: Computational and statistical perspectives. *Ann*. *Statist*. **51** 1–32. MR4564847 https://doi.org/10.1214/21-aos2155

HAZIMEH, H., MAZUMDER, R. and SAAB, A. (2022a). Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Math*. *Program*. **196** 347–388. MR4506808 https://doi.org/10.1007/s10107-021-01712-4

HAZIMEH, H., MAZUMDER, R. and SAAB, A. (2022b). Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Math*. *Program*. **196** 347–388. MR4506808 https://doi.org/10.1007/s10107-021-01712-4

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann*. *Statist*. **38** 2282–2313. MR2676890 https://doi.org/10.1214/09-AOS781

IBRAHIM, S., RADCHENKO, P., BEN-DAVID, E. and MAZUMDER, R. (2025). Supplement to “Predicting Census Survey Response Rates with Parsimonious Additive Models and Structured Interactions.” https://doi.org/10.1214/24-AOAS1929SUPPA, https://doi.org/10.1214/24-AOAS1929SUPPB

KINGMA, D. P. and BA, J. (2015). Adam: A method for stochastic optimization. In 3*rd International Conference on Learning Representations*, *ICLR* 2015, *San Diego*, *CA*, *USA*, *May* 7-9, 2015, *Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.).

KULZICK, R., KAIL, L., MULLENAX, S. et al. (2019). 2020 Census predictive models and audience segmentation report.

LI, K.-C., LUE, H.-H. and CHEN, C.-H. (2000). Interactive tree-structured regression via principal Hessian directions. *J*. *Amer*. *Statist*. *Assoc*. **95** 547–560. MR1803172 https://doi.org/10.2307/2669398

LIM, M. and HASTIE, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J*. *Comput*. *Graph*. *Statist*. **24** 627–654. MR3397226 https://doi.org/10.1080/10618600.2014.938812

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann*. *Statist*. **34** 2272–2297. MR2291500 https://doi.org/10.1214/009053606000000722

LOU, Y., CARUANA, R., GEHRKE, J. and HOOKER, G. (2013). Accurate intelligible models with pairwise interactions. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD* ’13 623–631. Association for Computing Machinery, New York, NY, USA.

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann*. *Statist*. **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann*. *Statist*. **27** 1443–1490. MR1742496 https://doi.org/10.1214/aos/1017939137

MAMMEN, E. and PARK, B. U. (2006). A simple smooth backfitting method for additive models. *Ann*. *Statist*. **34** 2252–2271. MR2291499 https://doi.org/10.1214/009053606000000696

MAZUMDER, R. and RADCHENKO, P. (2017). The *discrete Dantzig selector*: Estimating sparse linear models via mixed integer linear optimization. *IEEE Trans*. *Inf*. *Theory* **63** 3053–3075. MR3649777

MAZUMDER, R., RADCHENKO, P. and DEDIEU, A. (2023). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *Oper. Res.* **71** 129–147. MR4560192 https://doi.org/10.1287/opre.2022.2276

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443 https://doi.org/10.1214/09-AOS692

MULE, T. (2012). 2010 Census coverage measurement estimation report: Summary of estimates of coverage for persons in the United States DSSD 2010 Census Coverage Measurement Memorandum Series No. 2010-G-01 US Census Bureau.

NEMHAUSER, G. and WOLSEY, L. (1999). *Integer and Combinatorial Optimization. Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley, New York. MR1699321

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

NORI, H., JENKINS, S., KOCH, P. and CARUANA, R. (2019). InterpretML: A unified framework for machine learning interpretability. Available at arXiv:1909.09223.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

RADCHENKO, P. and JAMES, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Amer. Statist. Assoc.* **105** 1541–1553. MR2796570 https://doi.org/10.1198/jasa.2010.tm10130

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

SMITH, T. W. (2014). Hard-to-survey populations in comparative perspective. In *Hard-to-Survey Populations* 21–36. Cambridge Univ. Press, Cambridge.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606. MR0840516 https://doi.org/10.1214/aos/1176349940

TAN, Z. and ZHANG, C.-H. (2019). Doubly penalized estimation in additive regression with high-dimensional data. *Ann. Statist.* **47** 2567–2600. MR3988766 https://doi.org/10.1214/18-AOS1757

TOURANGEAU, R., EDWARDS, B., JOHNSON, T. P. et al. (2014). *Hard-to-Survey Populations*. Cambridge Univ. Press, Cambridge.

US CENSUS BUREAU (2009). 2010 Census integrated communications campaign plan: The success of the Census is in our hands.

US CENSUS BUREAU (2016). American community survey data suppression.

US CENSUS BUREAU (2017). Foundational communications models 1.0: The updated Low Response Score (LRS) and an Internet LRS.

US CENSUS BUREAU (2019). Planning database.

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

WOOD, S. N., LI, Z., SHADDICK, G. and AUGUSTIN, N. H. (2017). Generalized additive models for gigadata: Modeling the U.K. Black Smoke Network daily data. *J. Amer. Statist. Assoc.* **112** 1199–1210. MR3735370 https://doi.org/10.1080/01621459.2016.1195744

WRIGHT, S. J. (2015). Coordinate descent algorithms. *Math. Program.* **151** 3–34. MR3347548 https://doi.org/10.1007/s10107-015-0892-3

YAN, X. and BIEN, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statist. Sci.* **32** 531–560. MR3730521 https://doi.org/10.1214/17-STS622

YANG, Z., ZHANG, A. and SUDJIANTO, A. (2021). GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognit.* **120** 108192.

YUAN, M. and ZHOU, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.* **44** 2564–2593. MR3576554 https://doi.org/10.1214/15-AOS1422

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

ZHAO, T. and LIU, H. (2012). Sparse additive machine. In *AISTATS* (N. D. Lawrence and M. Girolami, eds.). *Proceedings of Machine Learning Research* **22** 1435–1443. PMLR, La Palma, Canary Islands.

ZSCHECH, P., WEINZIERL, S., HAMBAUER, N. et al. (2022). GAM(e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints. *ECIS* 2022 *Res. Pape.* **106**.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 121–146
https://doi.org/10.1214/24-AOAS1932

ULTRA-SPARSE SMALL AREA ESTIMATION WITH SUPER HEAVY-TAILED PRIORS FOR INTERNAL MIGRATION FLOWS

BY JAIRO FÚQUENE-PATIÑO[1,a] AND BRENDA BETANCOURT[2,b]

[1]*Department of Statistics, University of California, Davis,* [a]*jafuquenepatino@ucdavis.edu*
[2]*Department of Statistics and Data Science, NORC at the University of Chicago,* [b]*betancourt-brenda@norc.org*

Migration flows represent an important component of global sustainable development and demographic trends. However, the dynamic nature of the migration phenomenon, known issues of undercoverage of administrative records and long intercensal periods make estimation of internal migration a very challenging task. In this work we focus on the estimation of internal migration in Colombia, which is the subject of an ongoing armed conflict that has triggered forced and voluntary population movements from rural areas. Motivated by the high variability of migration flows across small areas in Colombia, we propose the use of super heavy-tailed (SHT) priors for sparse and ultra-sparse small area effects under a Fay–Herriot model.

We establish theoretical properties of a family of log-Cauchy priors and a new SHT prior obtained by considering a four parameter beta (FPB) density. We provide especially suited Markov chain Monte Carlo (MCMC) algorithms that can also be applied to other global-local families of priors in small area contexts. In addition, we consider a simulation study to illustrate how our proposal improves the precision of posterior estimates in sparse and ultra-sparse settings compared to other existing priors in the literature. Finally, we apply our proposed methodology to the estimation of internal migration in Colombia and obtain results with improved precision that are consistent with the population dynamics in the country. Moreover, we provide practical suggestions for official statisticians and other practitioners who desire to use our proposed framework in their own SAE problems.

REFERENCES

ANAND, S. and SEN, A. (1997). Concepts or human development and poverty! A multidimensional perspective. United Nations Development Programme, Poverty and Human Development: Human Development Papers 1–20.

ANGULO, R., DÍAZ, Y., PARDO, R. et al. (2013). Multidimensional poverty in Colombia, 1997–2010. Technical Report, Institute for Social and Economic Research.

ARMAGAN, A., DUNSON, D. B. and CLYDE, M. (2011). Generalized beta mixtures of Gaussians. *Adv. Neural Inf. Process. Syst.* **24** 523–531.

ARMAGAN, A., DUNSON, D. B. and LEE, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23** 119–143. MR3076161

AZOSE, J. J. and RAFTERY, A. E. (2018). Estimating large correlation matrices for international migration. *Ann. Appl. Stat.* **12** 940–970. MR3834291 https://doi.org/10.1214/18-AOAS1175

BAI, R. and GHOSH, M. (2019). Large-scale multiple hypothesis testing with the normal-beta prime prior. *Statistics* **53** 1210–1233. MR4034859 https://doi.org/10.1080/02331888.2019.1662017

BENAVENT, R. and MORALES, D. (2016). Multivariate Fay–Herriot models for small area estimation. *Comput. Statist. Data Anal.* **94** 372–390. MR3412832 https://doi.org/10.1016/j.csda.2015.07.013

BERGER, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 716–761. MR0572619

BHADRA, A., DATTA, J., POLSON, N. G. and WILLARD, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980 https://doi.org/10.1214/16-BA1028

BIJAK, J., BRYANT, J., GOŁATA, E. and SMALLWOOD, S. (2021). Preface. *J. Off. Stat.* **37** 533–541.

Key words and phrases. Small area estimation, internal migration, log-Cauchy priors, super heavy-tailed priors, sparsity, MCMC algorithms, Fay–Herriot model.

BISWAS, N., BHATTACHARYA, A., JACOB, P. E. and JOHNDROW, J. E. (2022). Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 973–996. MR4460582 https://doi.org/10.1111/rssb.12495

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

DANE (2008). Estimación e interpretación del coeficiente de variación de la encuesta cocensal.

DANE (2014). Guía para la Interpretación del Error Muestral en Términos del Coeficiente de Variación e Intervalo de Confianza Estimado. Encuesta de Sacrificio de Ganado.

DANE (2022). Reporte estadístico de migración número 2.

DATTA, G. S., GHOSH, M., STEORTS, R. and MAPLES, J. (2011). Bayesian benchmarking with applications to small area estimation. *TEST* **20** 574–588. MR2864715 https://doi.org/10.1007/s11749-010-0218-y

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* **10** 613–627. MR1769758

DATTA, G. S. and MANDAL, A. (2015). Small area estimation with uncertain random effects. *J. Amer. Statist. Assoc.* **110** 1735–1744. MR3449069 https://doi.org/10.1080/01621459.2015.1016526

DHS (2015). Colombia demographic and health survey. Technical Report, Ministerio de Salud y Protección Social y Profamilia.

DNP (2016). Índice de incidencia del conflicto armado. Departamento Nacional de Planeación—Grupo de Proyectos Especiales.

ECLAC (2013). Montevideo consensus on population and development. Technical Report, Economic Commission for Latin America and the Caribbean.

ECLAC (2019). First regional report on the implementation of the montevideo consensus on population and development. Technical Report, Economic Commission for Latin America and the Caribbean.

ECLAC AND UNFPA (2013). Principales cambios en las boletas de los censos latinoamericanos de las décadas de 1990, 2000 y 2010. Technical Report, Economic Commission for Latin America and the Caribbean and United Nations Population Fund.

EUROSTAT (2013). Handbook on precision requirements and variance estimation for ESS household surveys.

FALK, M., HÜSLER, J. and REISS, R.-D. (2011). Extreme value theory. In *Laws of Small Numbers: Extremes and Rare Events* 25–101. Springer, Berlin.

FAY, R. E. III 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

FRÜHWIRTH-SCHNATTER, S. and WAGNER, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In *Bayesian Statistics* 9 165–200. Oxford Univ. Press, Oxford. MR3204006 https://doi.org/10.1093/acprof:oso/9780199694587.003.0006

FÚQUENE-PATIÑO, J. and BETANCOURT, B. (2025). Supplement to "Ultra-sparse small area estimation with super heavy-tailed priors for internal migration flows." https://doi.org/10.1214/24-AOAS1932SUPPA, https://doi.org/10.1214/24-AOAS1932SUPPB

FÚQUENE-PATIÑO, J., CRISTANCHO, C., OSPINA, M. and GONZALEZ, D. M. (2021). Fay–Herriot model-based prediction alternatives for estimating households with emigrated members. *J. Off. Stat.* **37** 771–789.

GAGNON, P., DESGAGNÉ, A. and BÉDARD, M. (2020). A new Bayesian approach to robustness against outliers in linear regression. *Bayesian Anal.* **15** 389–414. MR4078719 https://doi.org/10.1214/19-BA1157

GAMMAGE, S. and STEVANOVIC, N. (2019). Gender, migration and care deficits: What role for the sustainable development goals? *J. Ethn. Migr. Stud.* **45** 2600–2620.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

RAO, J. et al. (2020). Discussion of "Small area estimation: Its evolution in five decades", by Malay Ghosh. *Stat. Trans. New Ser.* **21** 53–58.

GHOSH, P., TANG, X., GHOSH, M. and CHAKRABARTI, A. (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Anal.* **11** 753–796. MR3498045 https://doi.org/10.1214/15-BA973

GRIFFIN, J. and BROWN, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. University of Kent Technical Report.

HÁJEK, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". *Found. Surv. Sampl.* **236**.

HAMURA, Y., IRIE, K. and SUGASAWA, S. (2022a). On global-local shrinkage priors for count data. *Bayesian Anal.* **17** 545–564. MR4483230 https://doi.org/10.1214/21-ba1263

HAMURA, Y., IRIE, K. and SUGASAWA, S. (2022b). Log-regularly varying scale mixture of normals for robust regression. *Comput. Statist. Data Anal.* **173** Paper No. 107517. MR4418923 https://doi.org/10.1016/j.csda.2022.107517

AMNESTY INTERNATIONAL (2014). A land title is not enough: Ensuring sustainable land restitution in Colombia. Available at https://www.amnesty.org.

AMNESTY INTERNATIONAL (2019). Human rights in the Americas: Review of 2019. Available at https://www.amnesty.org.

IOM (2018). Migration and the 2030 agenda. Technical Report, Internationl Organization for Migration.

IOM (2021). Venezuelan refugee and migrant crisis. Technical Report, UN MIGRATION.

JIANG, J. and RAO, J. S. (2020). Robust small area estimation: An overview. *Annu. Rev. Stat. Appl.* **7** 337–360. MR4104196 https://doi.org/10.1146/annurev-statistics-031219-041212

JOHNDROW, J., ORENSTEIN, P. and BHATTACHARYA, A. (2020). Scalable approximate MCMC algorithms for the horseshoe prior. *J. Mach. Learn. Res.* **21** Paper No. 73. MR4095352

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135 https://doi.org/10.1214/009053604000000030

KLEIN SOLOMON, M. and SHELDON, S. (2018). The global compact for migration: From the sustainable development goals to a comprehensive agreement on safe, orderly and regular migration. *Int. J. Refug. Law*.

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

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

PLANE, D. A. (1982). An information theoretic approach to the estimation of migration flows. *J. Reg. Sci.* **22** 441–456.

POLSON, N. G. and SCOTT, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.* **7** 887–902. MR3000018 https://doi.org/10.1214/12-BA730

PRATESI, M., SALVATI, N. et al. (2009). Small area estimation in the presence of correlated random area effects. *J. Off. Stat.* **25** 37.

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

RAYMER, J. (2007). The estimation of international migration flows: A general technique focused on the origin-destination association structure. *Environ. Plan. A* **39** 985–995.

RAYMER, J. and ROGERS, A. (2007). Using age and spatial flow structures in the indirect estimation of migration streams. *Demography* **44** 199–223. https://doi.org/10.1353/dem.2007.0016

RESTREPO, J. A., SPAGAT, M. and VARGAS, J. F. (2003). The dynamics of the Colombian civil conflict: A new data set.

ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751 https://doi.org/10.1214/aoap/1034625254

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16** 351–367. MR1888450 https://doi.org/10.1214/ss/1015346320

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

SCHALL, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78** 719–727.

SDG (2018). Sustainable development goals. Technical Report, United Nations.

SHULTZ, J. M., CEBALLOS, Á. M. G., ESPINEL, Z., OLIVEROS, S. R., FONSECA, M. F. and FLOREZ, L. J. H. (2014). Internal displacement in Colombia: Fifteen distinguishing features. *Disaster Health* **2** 13–24. https://doi.org/10.4161/dish.27885

STATISTICS-CANADA (2014). Guide to the Labour Force Survey. Available at https://www150.statcan.gc.ca/n1/pub/71-543-g/71-543-g2014001-eng.pdf.

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Stat.* **42** 385–388. MR0397939 https://doi.org/10.1214/aoms/1177693528

TANG, X., GHOSH, M., HA, N. S. and SEDRANSK, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *J. Amer. Statist. Assoc.* **113** 1476–1489. MR3902223 https://doi.org/10.1080/01621459.2017.1419135

TANG, X., LI, K. and GHOSH, M. (2017). Bayesian multiple testing under sparsity for polynomial-tailed distributions. *Statist. Sinica* **27** 1225–1242. MR3699702

TZAVIDIS, N., ZHANG, L.-C., LUNA, A., SCHMID, T. and ROJAS-PERILLA, N. (2018). From start to finish: A framework for the production of small area official statistics. *J. Roy. Statist. Soc. Ser. A* **181** 927–979. MR3876378 https://doi.org/10.1111/rssa.12364

WILLEKENS, F. (2008). Models of migration: Observations and judgement. *Int. Migr. Eur: Data Mod. Est.* 117–147.

YANG, Z. (2019). Gaussian Regression Model Selection in High Dimension. PhD thesis, Indiana University.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 147–168
https://doi.org/10.1214/24-AOAS1946

POD-BIN: A PROBABILITY OF DECISION BAYESIAN INTERVAL DESIGN FOR TIME-TO-EVENT DOSE-FINDING TRIALS WITH MULTIPLE TOXICITY GRADES

BY MEIZI LIU[1,a], JI LIN[2,b], GU MI[2,c], CHRISTELLE LORENZATO[2,d], XUN CHEN[2,e] AND YUAN JI[3,f]

[1]*Statistical and Quantitative Sciences, Takeda,* [a]*meizi.liu@takeda.com*
[2]*Department of Biostatistics and Programming, Sanofi,* [b]*Ji.Lin@sanofi.com,* [c]*Gu.Mi@sanofi.com,* [d]*Christelle.Lorenzato@sanofi.com,* [e]*Xun.Chen@sanofi.com*
[3]*Department of Public Health Sciences, University of Chicago,* [f]*koaeraser@gmail.com*

We introduce a Bayesian framework centered on the "probability of decision" for designing dose-finding trials. The proposed PoD-BIN design evaluates the posterior predictive probabilities of up-and-down decisions. In PoD-BIN, multiple grades of toxicity, categorized as mild toxicity (MT) and dose-limiting toxicity (DLT), are simultaneously modeled, with the primary outcome being time-to-toxicity for both MT and DLT. This approach allows the enrollment of new patients while previously enrolled patients are still being monitored for toxicity, potentially reducing the trial duration. The Bayesian decision rules in PoD-BIN employ the probability of decisions to balance the trade-off between accelerating the trial and the risk of exposing patients to excessively toxic doses. Through numerical examples, we illustrate the trade-off between speed and safety of PoD-BIN and compare it with existing designs. PoD-BIN demonstrates the ability to control the frequency of risky decisions while simultaneously shortening trial duration in simulations.

REFERENCES

BEKELE, B. N. and THALL, P. F. (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J. Amer. Statist. Assoc.* **99** 26–35. MR2061885 https://doi.org/10.1198/016214504000000043

CHEUNG, Y. K. and CHAPPELL, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56** 1177–1182. MR1815616 https://doi.org/10.1111/j.0006-341X.2000.01177.x

EZZALFANI, M., ZOHAR, S., QIN, R., MANDREKAR, S. J. and LE DELEY, M.-C. (2013). Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Stat. Med.* **32** 2728–2746. MR3069903 https://doi.org/10.1002/sim.5737

FAKIH, M. G., KOPETZ, S., KUBOKI, Y., KIM, T. W., MUNSTER, P. N., KRAUSS, J. C., FALCHOOK, G. S., HAN, S.-W., HEINEMANN, V. et al. (2022). Sotorasib for previously treated colorectal cancers with KRASG12C mutation (CodeBreaK100): A prespecified analysis of a single-arm, phase 2 trial. *Lancet Oncol.* **23** 115–124.

GUO, W., JI, Y. and LI, D. (2019). R-TPI: Rolling toxicity probability interval design to shorten the duration and maintain safety of phase I trials. *J. Biopharm. Statist.* 1–14.

GUO, W., WANG, S.-J., YANG, S., LYNN, H. and JI, Y. (2017). A Bayesian interval dose-finding design addressing Ockham's razor: MTPI-2. *Contemp. Clin. Trials* **58** 23–33.

HONG, D. S., FAKIH, M. G., STRICKLER, J. H., DESAI, J., DURM, G. A., SHAPIRO, G. I., FALCHOOK, G. S., PRICE, T. J., SACHER, A. et al. (2020). KRAS. *N. Engl. J. Med.* **383** 1207–1217. https://doi.org/10.1056/NEJMoa1917239

IASONOS, A., ZOHAR, S. and O'QUIGLEY, J. (2011). Incorporating lower grade toxicity information into dose finding designs. *Clin. Trials* **8** 370–379. https://doi.org/10.1177/1740774511410732

JI, Y., LIU, P., LI, Y. and NEBIYOU BEKELE, B. (2010). A modified toxicity probability interval method for dose-finding trials. *Clin. Trials* **7** 653–663.

JOHNSON, V. E. and ALBERT, J. H. (1999). *Ordinal Data Modeling. Statistics for Social Science and Public Policy*. Springer, New York. MR1683018

Key words and phrases. Dose limiting toxicity, equivalence interval, latent probit regression, moderate toxicities, phase I trial, time to toxicity.

JR, G. W. S., TOI, M., PATRICK, N. et al. (2017). MONARCH 2: Abemaciclib in combination with fulvestrant in women with HR+/HER2-advanced breast cancer who had progressed while receiving endocrine therapy. *J. Clin. Oncol.* **35** 2875–2884.

LEE, S. M., CHENG, B. and CHEUNG, Y. K. (2011). Continual reassessment method with multiple toxicity constraints. *Biostatistics* **12** 386–398. https://doi.org/10.1093/biostatistics/kxq062

LEE, S. M., URSINO, M., CHEUNG, Y. K. and ZOHAR, S. (2019). Dose-finding designs for cumulative toxicities using multiple constraints. *Biostatistics* **20** 17–29. MR3892280 https://doi.org/10.1093/biostatistics/kxx059

LIU, M., LIN, J., MI, G., LORENZATO, C., CHEN, X. and JI, Y. (2025). Supplement to "PoD-BIN: a Probability of Decision Bayesian Interval Design for Time-to-Event Dose-Finding Trials with Multiple Toxicity Grades." https://doi.org/10.1214/24-AOAS1946SUPP

LIU, M., WANG, S.-J. and JI, Y. (2020). The i3+3 design for phase I clinical trials. *J. Biopharm. Statist.* **30** 294–304. https://doi.org/10.1080/10543406.2019.1636811

LIU, S., YIN, G. and YUAN, Y. (2013). Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *Ann. Appl. Stat.* **7** 2138–2156. MR3161716 https://doi.org/10.1214/13-AOAS661

MÜLLER, P., BERRY, D. A., GRIEVE, A. P. and KRAMS, M. (2006). A Bayesian decision-theoretic dose-finding trial. *Decis. Anal.* **3** 197–207.

NAKAJIMA, E. C., DREZNER, N., LI, X., MISHRA-KALYANI, P. S., LIU, Y., ZHAO, H., BI, Y., LIU, J., RAHMAN, A. et al. (2021). FDA approval summary: Sotorasib for KRAS G12C-mutated metastatic NSCLC. *Clin. Cancer Res.*

NATIONAL CANCER INSTITUTE (2021). Common Terminology Criteria for Adverse Events (CTCAE).

NEUENSCHWANDER, B., BRANSON, M. and GSPONER, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Stat. Med.* **27** 2420–2439. MR2432497 https://doi.org/10.1002/sim.3230

O'QUIGLEY, J., PEPE, M. and FISHER, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* **46** 33–48. MR1059105 https://doi.org/10.2307/2531628

PATNAIK, A., ROSEN, L. S., TOLANEY, S. M., TOLCHER, A. W., GOLDMAN, J. W., GANDHI, L., PAPADOPOULOS, K. P., BEERAM, M., RASCO, D. W. et al. (2016). Efficacy and safety of abemaciclib, an inhibitor of CDK4 and CDK6, for patients with breast cancer, non-small cell lung cancer, and other solid tumors. *Cancer Discov.* **6** 740–753. https://doi.org/10.1158/2159-8290.CD-16-0095

SHAH, M., RAHMAN, A., THEORET, M. R. and PAZDUR, R. (2021). The drug-dosing conundrum in oncology-when less is more. *N. Engl. J. Med.* **385** 1445–1447.

STORER, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45** 925–937. MR1029610 https://doi.org/10.2307/2531693

VAN METER, E. M., GARRETT-MAYER, E. and BANDYOPADHYAY, D. (2012). Dose-finding clinical trial design for ordinal toxicity grades using the continuation ratio model: An extension of the continual reassessment method. *Clin. Trials* **9** 303–313.

YAN, F., MANDREKAR, S. J. and YUAN, Y. (2017). Keyboard: A novel Bayesian toxicity probability interval design for phase I clinical trials. *Clin. Cancer Res.* **23** 3994–4003.

YUAN, Y., LIN, R., LI, D., NIE, L. and WARREN, K. E. (2018). Time-to-event Bayesian optimal interval design to accelerate phase I trials. *Clin. Cancer Res.* **24** 4921–4930. https://doi.org/10.1158/1078-0432.CCR-18-0246

YUAN, Y. and YIN, G. (2011). Robust EM continual reassessment method in oncology dose finding. *J. Amer. Statist. Assoc.* **106** 818–831. MR2894740 https://doi.org/10.1198/jasa.2011.ap09476

YUAN, Z., CHAPPELL, R. and BAILEY, H. (2007). The continual reassessment method for multiple toxicity grades: A Bayesian quasi-likelihood approach. *Biometrics* **63** 173–179, 313. MR2345586 https://doi.org/10.1111/j.1541-0420.2006.00666.x

ZHOU, T., GUO, W., JI, Y. et al. (2020). PoD-TPI: Probability-of-decision toxicity probability interval design to accelerate phase I trials. *Stat. Biosci.* **12** 124–145.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 169–190
https://doi.org/10.1214/24-AOAS1948

INFERRING SYNERGISTIC AND ANTAGONISTIC INTERACTIONS IN MIXTURES OF EXPOSURES

BY SHOUNAK CHATTOPADHYAY[1,a], STEPHANIE M. ENGEL[2,b] AND DAVID DUNSON[3,c]

[1]*Department of Biostatistics, University of California, Los Angeles,* [a]*shounak.chattopadhyay@ucla.edu*
[2]*Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill,* [b]*stephanie.engel@unc.edu*
[3]*Departments of Statistical Science & Mathematics, Duke University,* [c]*dunson@duke.edu*

There is abundant interest in assessing the joint effects of multiple exposures on human health. This is often referred to as the mixtures problem in environmental epidemiology and toxicology. Classically, studies have examined the adverse health effects of different chemicals one at a time, but there is concern that certain chemicals may act together to amplify each other's effects. Such amplification is referred to as *synergistic* interaction, while chemicals that inhibit each other's effects have *antagonistic* interactions. Current approaches for assessing the health effects of chemical mixtures do not explicitly consider synergy or antagonism in the modeling, instead focusing on either parametric or unconstrained nonparametric dose response surface modeling. The parametric case can be too inflexible, while nonparametric methods face a curse of dimensionality that leads to overly wiggly and uninterpretable surface estimates. We propose a Bayesian approach that decomposes the response surface into additive main effects and pairwise interaction effects and then detects synergistic and antagonistic interactions. Variable selection decisions for each interaction component are also provided. This Synergistic Antagonistic Interaction Detection (SAID) framework is evaluated relative to existing approaches using simulation experiments and an application to data from NHANES.

REFERENCES

ANTONELLI, J., MAZUMDAR, M., BELLINGER, D., CHRISTIANI, D., WRIGHT, R. and COULL, B. (2020). Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *Ann. Appl. Stat.* **14** 257–275. MR4085093 https://doi.org/10.1214/19-AOAS1307

BARR, D. B., WILDER, L. C., CAUDILL, S. P., GONZALEZ, A. J., NEEDHAM, L. L. and PIRKLE, J. L. (2005). Urinary creatinine concentrations in the U.S. population: Implications for urinary biologic monitoring measurements. *Environ. Health Perspect.* **113** 192–200. https://doi.org/10.1289/ehp.7337

BAXMANN, A. C., AHMED, M. S., MARQUES, N. C., MENON, V. B., PEREIRA, A. B., KIRSZTAJN, G. M. and HEILBERG, I. P. (2008). Influence of muscle mass and physical activity on serum and urinary creatinine and serum cystatin C. *Clin. J. Amer. Soc. Nephrol.* **3** 348–354. https://doi.org/10.2215/CJN.02870707

BETANCOURT, M. and GIROLAMI, M. (2015). Hamiltonian Monte Carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications* 79–101. CRC Press, Boca Raton, FL. MR3644666

BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2012). Bayesian shrinkage. ArXiv preprint. Available at arXiv:1212.6088.

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. MR3365442 https://doi.org/10.1093/biostatistics/kxu058

BREZGER, A. and LANG, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Comput. Statist. Data Anal.* **50** 967–991. MR2210741 https://doi.org/10.1016/j.csda.2004.10.011

Key words and phrases. Additive model, Bayesian, dose response surface, nonparametric regression, pairwise interaction, positivity constraint, synergy, exposure mixtures.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80. PMLR.

CHATTOPADHYAY, S., ENGEL, S. M. and DUNSON, D. (2025). Supplement to "Inferring synergistic and antagonistic interactions in mixtures of exposures." https://doi.org/10.1214/24-AOAS1948SUPPA, https://doi.org/10.1214/24-AOAS1948SUPPB

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

CZARNOTA, J., GENNINGS, C. and WHEELER, D. C. (2015). Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. *Cancer Inform.* **14** 159–171. https://doi.org/10.4137/CIN.S17295

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

FERRARI, F. and DUNSON, D. B. (2020). Identifying main effects and interactions among exposures using Gaussian processes. *Ann. Appl. Stat.* **14** 1743–1758. MR4194246 https://doi.org/10.1214/20-AOAS1363

FERRARI, F. and DUNSON, D. B. (2021). Bayesian factor analysis for inference on interactions. *J. Amer. Statist. Assoc.* **116** 1521–1532. MR4309290 https://doi.org/10.1080/01621459.2020.1745813

FERRARO, P. M., COSTANZI, S., NATICCHIA, A., STURNIOLO, A. and GAMBARO, G. (2010). Low level exposure to cadmium increases the risk of chronic kidney disease: Analysis of the NHANES 1999-2006. *BMC Public Health* **10** 1–8.

FORBES, G. B. and BRUINING, G. J. (1976). Urinary creatinine excretion and lean body mass. *Amer. J. Clin. Nutr.* **29** 1359–1366. https://doi.org/10.1093/ajcn/29.12.1359

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. With comments and a rejoinder by the authors. MR1422404

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

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

HAO, N., FENG, Y. and ZHANG, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *J. Amer. Statist. Assoc.* **113** 615–625. MR3832213 https://doi.org/10.1080/01621459.2016.1264956

HARIS, A., WITTEN, D. and SIMON, N. (2016). Convex modeling of interactions with strong heredity. *J. Comput. Graph. Statist.* **25** 981–1004. MR3572025 https://doi.org/10.1080/10618600.2015.1067217

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

HAYS, S. M., AYLWARD, L. L. and BLOUNT, B. C. (2015). Variation in urinary flow rates according to demographic characteristics and body mass index in NHANES: Potential confounding of associations between health outcomes and urinary biomarker concentrations. *Environ. Health Perspect.* **123** 293–300. https://doi.org/10.1289/ehp.1408944

HERRING, A. H. (2010). Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection. *Epidemiology* **21** S71.

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

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

JAIN, R. (2016). Associated complex of urine creatinine, serum creatinine, and chronic kidney disease. *Epidemiology* **6** 2161–1165.

JAMES, G. D., SEALEY, J. E., ALDERMAN, M., LJUNGMAN, S., MUELLER, F. B., PECKER, M. S. and LARAGH, J. H. (1988). A longitudinal study of urinary creatinine and creatinine clearance in normal subjects. Race, sex, and age differences. *J. Hypertens.* **1** 124–131. https://doi.org/10.1093/ajh/1.2.124

JENG, P.-H., HUANG, T.-R., WANG, C.-C. and CHEN, W.-L. (2021). Clinical relevance of urine flow rate and exposure to polycyclic aromatic hydrocarbons. *Int. J. Environ. Res. Public Health* **18** 5372.

JOUBERT, B. R., KIOUMOURTZOGLOU, M.-A., CHAMBERLAIN, T., CHEN, H. Y., GENNINGS, C., TURYK, M. E., MIRANDA, M. L., WEBSTER, T. F., ENSOR, K. B. et al. (2022). Powering Research through Innovative Methods for Mixtures in Epidemiology (PRIME) program: Novel and expanded statistical methods. *Int. J. Environ. Res. Public Health* **19** 1378.

KASHANI, K., ROSNER, M. H. and OSTERMANN, M. (2020). Creatinine: From physiology to clinical application. *Eur. J. Intern. Med.* **72** 9–14. https://doi.org/10.1016/j.ejim.2019.10.025

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

KIM, N. H., HYUN, Y. Y., LEE, K.-B., CHANG, Y., RHU, S., OH, K.-H. and AHN, C. (2015). Environmental heavy metal exposure and chronic kidney disease in the general population. *J. Korean Med. Sci.* **30** 272–277.

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

LI, H. and PATI, D. (2017). Variable selection using shrinkage priors. *Comput. Statist. Data Anal.* **107** 107–119. MR3575062 https://doi.org/10.1016/j.csda.2016.10.008

LUO, J. and HENDRYX, M. (2020). Metal mixtures and kidney function: An application of machine learning to NHANES data. *Environ. Res.* **191** 110126. https://doi.org/10.1016/j.envres.2020.110126

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

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

MIDDLETON, D. R., WATTS, M. J., LARK, R. M., MILNE, C. J. and POLYA, D. A. (2016). Assessing urinary flow rate, creatinine, osmolality and other hydration adjustment methods for urinary biomonitoring using NHANES arsenic, iodine, lead and cadmium data. *Environ. Health* **15** 1–13.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. With comments by James Berger and C. L. Mallows and with a reply by the authors. MR0997578

MOLITOR, J., PAPATHOMAS, M., JERRETT, M. and RICHARDSON, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics* **11** 484–498.

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

POLLACK, A. Z., MUMFORD, S. L., MENDOLA, P., PERKINS, N. J., ROTMAN, Y., WACTAWSKI-WENDE, J. and SCHISTERMAN, E. F. (2015). Kidney biomarkers associated with blood lead, mercury, and cadmium in premenopausal women: A prospective cohort study. *J. Toxicol. Environ. Health* **78** 119–131. https://doi.org/10.1080/15287394.2014.944680

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

RAHMAN, H. H., NIEMANN, D. and MUNSON-MCGEE, S. H. (2022). Association of albumin to creatinine ratio with urinary arsenic and metal exposure: Evidence from NHANES 2015-2016. *Int. Urol. Nephrol.* **54** 1343–1353. https://doi.org/10.1007/s11255-021-03018-y

RUE, H. (2001). Fast sampling of Gaussian Markov random fields. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 325–338. MR1841418 https://doi.org/10.1111/1467-9868.00288

SAMANTA, S. and ANTONELLI, J. (2022). Estimation and false discovery control for the analysis of environmental mixtures. *Biostatistics* **23** 1039–1055. MR4496361 https://doi.org/10.1093/biostatistics/kxac001

WANG, C., JIANG, B. and ZHU, L. (2021). Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statist. Sinica* **31** 1549–1570. MR4297716 https://doi.org/10.5705/ss.202019.0081

WEI, R., REICH, B. J., HOPPIN, J. A. and GHOSAL, S. (2020). Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statist. Sinica* **30** 55–79. MR4285485 https://doi.org/10.5705/ss.202017.0315

WILLIAMS, C. K. I. and RASMUSSEN, C. E. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 191–211
https://doi.org/10.1214/24-AOAS1951

CONTOUR LOCATION FOR RELIABILITY IN AIRFOIL SIMULATION EXPERIMENTS USING DEEP GAUSSIAN PROCESSES

BY ANNIE S. BOOTH[1,a], S. ASHWIN RENGANATHAN[2,b] AND ROBERT B. GRAMACY[3,c]

[1]*Department of Statistics, North Carolina State University,* [a]*annie_booth@ncsu.edu*
[2]*Department of Aerospace Engineering, Pennsylvania State University,* [b]*ashwin.renganathan@psu.edu*
[3]*Department of Statistics, Virginia Tech,* [c]*rbg@vt.edu*

Bayesian deep Gaussian processes (DGPs) outperform ordinary GPs as surrogate models of complex computer experiments when response surface dynamics are nonstationary, which is especially prevalent in aerospace simulations. Yet DGP surrogates have not been deployed for the canonical downstream task in that setting: reliability analysis through contour location (CL). In that context we are motivated by a simulation of an RAE-2822 transonic airfoil which demarcates efficient and inefficient flight conditions. Level sets separating passable vs. failable operating conditions are best learned through strategic sequential designs. There are two limitations to modern CL methodology which hinder DGP integration in this setting. First, derivative-based optimization underlying acquisition functions is thwarted by sampling-based Bayesian (i.e., MCMC) inference, which is essential for DGP posterior integration. Second, canonical acquisition criteria, such as entropy, are famously myopic to the extent that optimization may even be undesirable. Here we tackle both of these limitations at once, proposing a hybrid criterion that explores along the Pareto front of entropy and (predictive) uncertainty, requiring evaluation only at strategically located "triangulation" candidates. We showcase DGP CL performance in several synthetic benchmark exercises and on the RAE-2822 airfoil.

REFERENCES

ABDELMALEK-LOMENECH, R. A., BECT, J., CHABRIDON, V. and VAZQUEZ, E. (2022). Bayesian sequential design of computer experiments to estimate reliable sets. Available at arXiv:2211.01008.

ALLEN, M. and MAUTE, K. (2004). Reliability-based design optimization of aeroelastic structures. *Struct. Multidiscip. Optim.* **27** 228–242.

BARBER, C. B., DOBKIN, D. P. and HUHDANPAA, H. (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Software* **22** 469–483. MR1428265 https://doi.org/10.1145/235815.235821

BASTOS, L. S. and O'HAGAN, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51** 425–438. MR2756478 https://doi.org/10.1198/TECH.2009.08019

BICHON, B. J., ELDRED, M. S., SWILER, L. P., MAHADEVAN, S. and MCFARLAND, J. M. (2008). Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA J.* **46** 2459–2468.

BINOIS, M., GINSBOURGER, D. and ROUSTANT, O. (2015). Quantifying uncertainty on Pareto fronts with Gaussian process conditional simulations. *European J. Oper. Res.* **243** 386–394. MR3315549 https://doi.org/10.1016/j.ejor.2014.07.032

BITZER, M., MEISTER, M. and ZIMMER, C. (2023). Hierarchical-hyperplane kernels for actively earning Gaussian process models of nonstationary systems. In *International Conference on Artificial Intelligence and Statistics* 7897–7912. PMLR, Valencia, Spain.

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

BOOTH, A. S. (2023). deepgp: Deep Gaussian processes using MCMC. R package version 1.1.2.

BOOTH, A. S., RENGANATHAN, S. A. and GRAMACY, R. B. (2025). Supplement to "Contour location for reliability in airfoil simulation experiments using deep Gaussian processes." https://doi.org/10.1214/24-AOAS1951SUPPA, https://doi.org/10.1214/24-AOAS1951SUPPB

Key words and phrases. Active learning, emulator, entropy, Pareto front, sequential design, surrogate, tricands, uncertainty quantification.

BREMER, S. A. (2019). *The GLOBUS Model: Computer Simulation of Worldwide Political and Economic Developments*. Routledge, London.

BRYAN, B., NICHOL, R. C., GENOVESE, C. R., SCHNEIDER, J., MILLER, C. J. and WASSERMAN, L. (2005). Active learning for identifying function threshold boundaries. *Adv. Neural Inf. Process. Syst.* **18**.

BUI, T., HERNÁNDEZ-LOBATO, D., HERNANDEZ-LOBATO, J., LI, Y. and TURNER, R. (2016). Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning* 1472–1481. PMLR, New York, USA.

CHEVALIER, C., GINSBOURGER, D., BECT, J., VAZQUEZ, E., PICHENY, V. and RICHET, Y. (2014). Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics* **56** 455–465. MR3290615 https://doi.org/10.1080/00401706.2013.860918

COHN, D. (1994). Neural network exploration using optimal experiment design. *Adv. Neural Inf. Process. Syst.* **6** 679–686.

COLE, D. A., GRAMACY, R. B., WARNER, J. E., BOMARITO, G. F., LESER, P. E. and LESER, W. P. (2022). Entropy-based adaptive design for contour finding and estimating reliability. *J. Qual. Technol.* 1–18.

DAMIANOU, A. and LAWRENCE, N. D. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics* 207–215. PMLR, Arizona, USA.

DUNLOP, M. M., GIROLAMI, M. A., STUART, A. M. and TECKENTRUP, A. L. (2018). How deep are deep Gaussian processes? *J. Mach. Learn. Res.* **19** Paper No. 54, 46. MR3874162

ECONOMON, T. D., PALACIOS, F., COPELAND, S. R., LUKACZYK, T. W. and ALONSO, J. J. (2016). SU2: An open-source suite for multiphysics simulation and design. *AIAA J.* **54** 828–846.

ERIKSSON, D., PEARCE, M., GARDNER, J., TURNER, R. D. and POLOCZEK, M. (2019). Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems* **32**. Curran Associates, Red Hook.

GARDNER, J., PLEISS, G., WEINBERGER, K. Q., BINDEL, D. and WILSON, A. G. (2018). Gpytorch: Blackbox matrix–matrix Gaussian process inference with gpu acceleration. *Adv. Neural Inf. Process. Syst.* **31**.

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

GOODARZI, E., ZIAEI, M. and HOSSEINIPOUR, E. Z. (2014). *Introduction to Optimization Analysis in Hydrosystem Engineering*. Springer, Cham.

GRAMACY, R. B. (2020). *Surrogates—Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Chapman & Hall/CRC Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR4283556 https://doi.org/10.1201/9780367815493

GRAMACY, R. B. and APLEY, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.* **24** 561–578. MR3357395 https://doi.org/10.1080/10618600.2014.914442

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

GRAMACY, R. B. and LEE, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145. MR2668170 https://doi.org/10.1198/TECH.2009.0015

GRAMACY, R. B., SAUER, A. and WYCOFF, N. (2022). Triangulation candidates for Bayesian optimization. *Adv. Neural Inf. Process. Syst.* **35** 35933–35945.

GRAY, R. M. (2011). *Entropy and Information Theory*, 2nd ed. Springer, New York. MR3134681 https://doi.org/10.1007/978-1-4419-7970-4

HABEL, K., GRASMAN, R., GRAMACY, R. B., MOZHAROVSKYI, P. and STERRATT, D. C. (2019). geometry: Mesh generation and surface tessellation. R package version 0.4.5.

HIGDON, D., SWALL, J. and KERN, J. (1999). Non-stationary spatial modeling. *Bayesian Stat.* **6** 761–768.

IZZATURRAHMAN, M. F., PALAR, P. S., ZUHAL, L. and SHIMOYAMA, K. (2022). Modeling non-stationarity with deep Gaussian processes: Applications in aerospace engineering. In *AIAA Scitech Forum* 1096.

JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *J. Global Optim.* **13** 455–492. MR1673460 https://doi.org/10.1023/A:1008306431147

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

LIPPE, M., BITHELL, M., GOTTS, N., NATALINI, D., BARBROOK-JOHNSON, P., GIUPPONI, C., HALLIER, M., HOFSTEDE, G. J., LE PAGE, C. et al. (2019). Using agent-based modelling to simulate social-ecological systems across scales. *GeoInformatica* **23** 269–298.

LUO, J., GUPTA, A., ONG, Y.-S. and WANG, Z. (2018). Evolutionary optimization of expensive multiobjective problems with co-sub-Pareto front Gaussian process surrogates. *IEEE Trans. Cybern.* **49** 1708–1721.

MARQUES, A., LAM, R. and WILLCOX, K. (2018). Contour location via entropy reduction leveraging multiple information sources. *Adv. Neural Inf. Process. Syst.* **31**.

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

Mehta, P. M., Walker, A., Lawrence, E., Linares, R., Higdon, D. and Koller, J. (2014). Modeling satellite drag coefficients with response surfaces. *Adv. Space Res.* **54** 1590–1607.

Ming, D., Williamson, D. and Guillas, S. (2023). Deep Gaussian process emulation using stochastic imputation. *Technometrics* **65** 150–161. MR4580865 https://doi.org/10.1080/00401706.2022.2124311

Murray, I., Adams, R. P. and MacKay, D. J. C. (2010). Elliptical slice sampling. In *The Proceedings of the* 13*th International Conference on Artificial Intelligence and Statistics*. *JMLR*: *W&CP* **9** 541–548. PMLR, Sardinia, Italy.

Oakley, J. (2004). Estimating percentiles of uncertain computer code outputs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 83–93. MR2043762 https://doi.org/10.1046/j.0035-9254.2003.05044.x

Paciorek, C. J. and Schervish, M. J. (2003). Nonstationary covariance functions for Gaussian process regression. In *Proceedings of the* 16*th International Conference on Neural Information Processing Systems*. *NIPS'*03 273–280. MIT Press, Cambridge, MA, USA.

Pamadi, B., Covell, P., Tartabini, P. and Murphy, K. (2004). Aerodynamic characteristics and glide-back performance of langley glide-back booster. In 22*nd Applied Aerodynamics Conference and Exhibit* 5382.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12** 2825–2830. MR2854348

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T. and Kim, N.-H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *J. Mech. Des.* **132**.

Radaideh, M. I. and Kozlowski, T. (2020). Surrogate modeling of advanced computer simulations using deep Gaussian processes. *Reliab. Eng. Syst. Saf.* **195** 106731.

Ranjan, P., Bingham, D. and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics* **50** 527–541. MR2655651 https://doi.org/10.1198/004017008000000541

Renganathan, A., Rao, V. and Navon, I. (2022). Multifidelity Gaussian processes for failure boundary and probability estimation. In *AIAA Scitech Forum* 0390.

Renganathan, S. A., Larson, J. and Wild, S. M. (2021). Lookahead acquisition functions for finite-horizon time-dependent Bayesian optimization and application to quantum optimal control. Available at arXiv:2105.09824.

Renganathan, S. A., Rao, V. and Navon, I. M. (2023). CAMERA: A method for cost-aware, adaptive, multifidelity, efficient reliability analysis. *J. Comput. Phys.* **472** Paper No. 111698, 25. MR4502218 https://doi.org/10.1016/j.jcp.2022.111698

Roocks, P. (2016). Computing Pareto frontiers and database preferences with the rPref package. *R J.* **8** 393–404.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. Available at arXiv:1705.08933.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.

Santner, T. J., Williams, B. J. and Notz, W. I. (2018). *The Design and Analysis of Computer Experiments*. *Springer Series in Statistics*. Springer, New York. 2nd ed. of [MR2160708]. MR3887662

Sauer, A., Cooper, A. and Gramacy, R. B. (2023a). Vecchia-approximated deep Gaussian processes for computer experiments. *J. Comput. Graph. Statist.* **32** 824–837. MR4641462 https://doi.org/10.1080/10618600.2022.2129662

Sauer, A., Cooper, A. and Gramacy, R. B. (2023b). Non-stationary Gaussian process surrogates. Available at arXiv:2305.19242.

Sauer, A., Gramacy, R. B. and Higdon, D. (2023). Active learning for deep Gaussian process surrogates. *Technometrics* **65** 4–18. MR4543056 https://doi.org/10.1080/00401706.2021.2008505

Sauer, A. E. (2023). Deep Gaussian process surrogates for computer experiments. Ph.D. thesis, Virginia Tech.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 743–758. MR1998632 https://doi.org/10.1111/1467-9868.00413

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Adv. Neural Inf. Process. Syst.* **18** 1259–1266.

Stanford, B., Sauer, A., Jacobson, K. and Warner, J. (2022). Gradient-enhanced reliability analysis of transonic aeroelastic flutter. In *AIAA Scitech Forum* 0632.

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

SURJANOVIC, S. and BINGHAM, D. (2013). Virtual library of simulation experiments: Test functions and datasets. Available at http://www.sfu.ca/~ssurjano.

TIETZE, N. (2015). Model-based calibration of engine control units using Gaussian process regression. Ph.D. thesis, Technische Universität.

VASSBERG, J., DEHAAN, M., RIVERS, M. and WAHLS, R. (2008). Development of a common research model for applied CFD validation studies. In 26*th AIAA Applied Aerodynamics Conference* 6919.

ZIMMER, L., LINDAUER, M. and HUTTER, F. (2021). Auto-pytorch: Multi-fidelity MetaLearning for efficient and robust AutoDL. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **43** 3079–3090. https://doi.org/10.1109/TPAMI.2021.3067763

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 212–234
https://doi.org/10.1214/24-AOAS1956

HAS THE COVID-19 OUTBREAK CAPSIZED THE PREDICTIVE PERFORMANCE OF BAYESIAN VAR MODELS WITH COINTEGRATION AND TIME-VARYING VOLATILITY?

BY ANNA PAJOR[1,a], ŁUKASZ KWIATKOWSKI[2,b] AND JUSTYNA WRÓBLEWSKA[2,c]

[1]*Department of Mathematics, Krakow University of Economics,* [a]*pajora@uek.krakow.pl*
[2]*Department of Econometrics and Operations Research, Krakow University or Economics,* [b]*kwiatkol@uek.krakow.pl,* [c]*wroblewj@uek.krakow.pl*

We check whether taking into account long-term relationships in heteroscedastic VAR models affects their predictive performance before and during the Covid-19 pandemic. Also, we examine whether the predictions can benefit from suspending posterior updates at some point. Empirical analysis covers five different economies and uses Bayesian VAR/VEC models with volatility specifications combining stochastic volatility and GARCH processes. It emerges that, while accounting for cointegration relationships in the models enhances their predictive performance prior to the pandemic, it may be counterproductive for times of economic crisis. Additionally, refraining from keeping the posterior updated does improve the predictions, but only rarely.

REFERENCES

APERGIS, N., COORAY, A., KHRAIEF, N. and APERGIS, I. (2019). Do gold prices respond to real interest rates? Evidence from the Bayesian Markov switching VECM model. *J. Int. Financ. Mark. Inst. Money* **60** 134–148.

BAUWENS, L., KOOP, G., KOROBILIS, D. and ROMBOUTS, J. V. K. (2015). The contribution of structural break models to forecasting macroeconomic series. *J. Appl. Econometrics* **30** 596–620. MR3358638 https://doi.org/10.1002/jae.2387

BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, Chichester. MR1274699 https://doi.org/10.1002/9780470316870

BOBEICA, E. and HARTWIG, B. (2023). The COVID-19 shock and challenges for inflation modelling. *Int. J. Forecast.* **39** 519–539. https://doi.org/10.1016/j.ijforecast.2022.01.002

CARRIERO, A., CLARK, T. E. and MARCELLINO, M. (2022). Nowcasting tail risk to economic activity at a weekly frequency. *J. Appl. Econometrics* **37** 843–866. MR4470806

CARRIERO, A., CLARK, T. E., MARCELLINO, M. and MERTENS, E. (2024). Addressing COVID-19 outliers in BVARs with stochastic volatility. *Rev. Econ. Stat.* **106** 1403–1417. https://doi.org/10.1162/rest_a_01213

CHAN, J. C. C. (2020). Large Bayesian VARs: A flexible Kronecker error covariance structure. *J. Bus. Econom. Statist.* **38** 68–79. MR4044952 https://doi.org/10.1080/07350015.2018.1451336

CHIKUSE, Y. (2003). *Statistics on Special Manifolds. Lecture Notes in Statistics* **174**. Springer, New York. MR1960435 https://doi.org/10.1007/978-0-387-21540-2

CHIU, C.-W. J., MUMTAZ, H. and PINTER, G. (2017). Forecasting with VAR models: Fat tails and stochastic volatility. *Int. J. Forecast.* **33** 1124–1143.

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

CUBADDA, G. (1999). Common cycles in seasonal non-stationary time series. *J. Appl. Econometrics* **14** 273–291.

ENGLE, R. F. and KRONER, K. F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory* **11** 122–150. MR1325104 https://doi.org/10.1017/S0266466600009063

FERRARA, L., MOGLIANI, M. and SAHUC, J.-G. (2022). High-frequency monitoring of growth at risk. *Int. J. Forecast.* **38** 582–595.

Key words and phrases. Bayesian forecasting, cointegrated VAR, GARCH models, stochastic volatility, SV-GARCH models, VEC models.

FORONI, C., MARCELLINO, M. and STEVANOVIC, D. (2022). Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis. *Int. J. Forecast.* **38** 596–612. https://doi.org/10.1016/j.ijforecast.2020.12.005

GANICS, G. and ODENDAHL, F. (2021). Bayesian VAR forecasts, survey information, and structural change in the euro area. *Int. J. Forecast.* **37** 971–999.

GEWEKE, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2166278 https://doi.org/10.1002/0471744735

GIANNONE, D., LENZA, M. and PRIMICERI, G. E. (2019). Priors for the long run. *J. Amer. Statist. Assoc.* **114** 565–580. MR3963163 https://doi.org/10.1080/01621459.2018.1483826

GRANGER, C. W. J. and SIKLOS, P. L. (1995). Systematic sampling, temporal aggregation, seasonal adjustment, and cointegration theory and evidence. *J. Econometrics* **66** 357–369.

HAMILTON, J. D. (2018). Why you should never use the Hodrick-Prescott filter. *Rev. Econ. Stat.* **100** 831–843.

HARTWIG, B. (2024). Bayesian VARs and prior calibration in times of COVID-19. *Stud. Nonlinear Dyn. Econom.* **28** 1–24. MR4713978 https://doi.org/10.1515/snde-2021-0108

HAUZENBERGER, N., HUBER, F., PFARRHOFER, M. and ZÖRNER, T. O. (2021). Stochastic model specification in Markov switching vector error correction models. *Stud. Nonlinear Dyn. Econom.* **25** 20180069. MR4255821 https://doi.org/10.1515/snde-2018-0069

HECQ, A. (1998). Does seasonal adjustment induce common cycles? *Econom. Lett.* **59** 289–297.

HUBER, F., KOOP, G., ONORANTE, L., PFARRHOFER, M. and SCHREINER, J. (2023). Nowcasting in a pandemic using non-parametric mixed frequency VARs. *J. Econometrics* **232** 52–69. MR4513100 https://doi.org/10.1016/j.jeconom.2020.11.006

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

JACQUIER, E., POLSON, N. G. and ROSSI, P. E. (1995). Models and priors for multivariate stochastic volatility. CIRANO Working Papers No. 95s-18, CIRANO.

JOCHMANN, M. and KOOP, G. (2015). Regime-switching cointegration. *Stud. Nonlinear Dyn. Econom.* **19** 35–48. MR3315033 https://doi.org/10.1515/snde-2012-0064

JUSELIUS, K. (2006). *The Cointegrated VAR Model*: *Methodology and Applications*. *Advanced Texts in Econometrics*. Oxford Univ. Press, Oxford. MR2358946

KOOP, G., LEÓN-GONZÁLEZ, R. and STRACHAN, R. W. (2009). Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Rev.* **29** 224–242. MR2747499 https://doi.org/10.1080/07474930903382208

KOOP, G., MCINTYRE, S., MITCHELL, J. and POON, A. (2021). Nowcasting 'true' monthly U.S. GDP during the pandemic. *Nat. Instit. Econ. Rev.* **256** 44–70.

KOOP, G., MCINTYRE, S., MITCHELL, J. and POON, A. (2023). Reconciled estimates of monthly GDP in the United States. *J. Bus. Econom. Statist.* **41** 563–577. MR4568043 https://doi.org/10.1080/07350015.2022.2044336

KUNST, R. M. and FRANSES, P. H. (1998). The impact of seasonal constants on forecasting seasonally cointegrated time series. *J. Forecast.* **17** 109–124.

KWIATKOWSKI, Ł. (2020a). Bayesian VEC models with Markov-switching heteroscedasticity in forecasting macroeconomic time series. In *The* 14*th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena* (M. Papież and S. Śmiech, eds.) 78–87. Wydawnictwo Małopolskiej Szkoły Administracji Publicznej Uniwersytetu Ekonomicznego w Krakowie.

KWIATKOWSKI, Ł. (2020b). Predictive power comparison of Bayesian homoscedastic vs. Markov-switching heteroscedastic VEC models. In *Quantitative Methods in the Contemporary Issues of Economics* (B. Ciałowicz, ed.) 67–76. edu-Libri s.c., Kraków—Legionowo.

LENZA, M. and PRIMICERI, G. E. (2022). How to estimate a vector autoregression after March 2020. *J. Appl. Econometrics* **37** 688–699. MR4450598 https://doi.org/10.1002/jae.2895

LÖF, M. and FRANSES, P. H. (2001). On forecasting cointegrated seasonal time series. *Int. J. Forecast.* **17** 607–621.

LÜTKEPOHL, H. and NETŠUNAJEV, A. (2018). The relation between monetary policy and the stock market in Europe. *Econometrics* **6**. https://doi.org/10.3390/econometrics6030036

MEYER, M. and WINKER, P. (2005). Using HP filtered data for econometric analysis: Some evidence from Monte Carlo simulations. *Allg. Stat. Arch.* **89** 303–320. MR2196143 https://doi.org/10.1007/s10182-005-0206-9

NERLOVE, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica* **32** 241–286.

OSIEWALSKI, J. (2009). New hybrid models of multivariate volatility (a Bayesian perspective). *Przeglad Statystyczny* **56** 15–22.

OSIEWALSKI, J. and PAJOR, A. (2009). Bayesian analysis for hybrid MSF-SBEKK models of multivariate volatility. *Central Eur. J. Econ. Modell. Econom.* **1** 179–202.

OSIEWALSKI, J. and PAJOR, A. (2018). A hybrid MSV-MGARCH generalisation of the t-MGARCH model. In *The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena*: *Conference Proceedings* (M. Papież and S. Śmiech, eds.) 345–354. Foundation of the Cracow Univ. Economics.

OSIEWALSKI, J. and PAJOR, A. (2019). On sensitivity of inference in Bayesian MSF-MGARCH models. *Central Eur. J. Econ. Modell. Econom.* **3** 137–197.

PAJOR, A. (2011). A Bayesian analysis of exogeneity in models with latent variables. *Central Eur. J. Econ. Modell. Econom.* **3** 49–73.

PAJOR, A. (2020). MCMC method for the IG-MSF-SBEKK model. In *Quantitative Methods in the Contemporary Issues of Economics* (B. Ciałowicz, ed.) 77–89. edu-Libri s.c., Kraków—Legionowo.

PAJOR, A., KWIATKOWSKI, Ł. and WRÓBLEWSKA, J. (2025a). Supplement to "Has the Covid-19 outbreak capsized the predictive performance of Bayesian VAR models with cointegration and time-varying volatility?." https://doi.org/10.1214/24-AOAS1956SUPPA

PAJOR, A., KWIATKOWSKI, Ł. and WRÓBLEWSKA, J. (2025b). Supplement to "Has the Covid-19 outbreak capsized the predictive performance of Bayesian VAR models with cointegration and time-varying volatility?." https://doi.org/10.1214/24-AOAS1956SUPPB

PAJOR, A., OSIEWALSKI, J., WRÓBLEWSKA, J. and KWIATKOWSKI, Ł. (2024b). Bayesian ex post evaluation of recursive multi-step-ahead density prediction. *Bayesian Anal.* **19** 751–783. MR4769123 https://doi.org/10.1214/23-ba1367

PAJOR, A. and WRÓBLEWSKA, J. (2017). VEC-MSF models in Bayesian analysis of short- and long-run relationships. *Stud. Nonlinear Dyn. Econom.* **21** 20160004. MR3665986 https://doi.org/10.1515/snde-2016-0004

PAJOR, A. and WRÓBLEWSKA, J. (2022). Forecasting performance of Bayesian VEC-MSF models for financial data in the presence of long-run relationships. *Eurasian Econ. Rev.* **12** 427–448.

PAJOR, A., WRÓBLEWSKA, J., KWIATKOWSKI, Ł. and OSIEWALSKI, J. (2024a). Hybrid SV-GARCH, t-GARCH and Markov-switching covariance structures in VEC models—which is better from a predictive perspective? *Int. Stat. Rev.* **92** 62–86. MR4727440 https://doi.org/10.1111/insr.12546

PRIMICERI, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Rev. Econ. Stud.* **72** 821–852. MR2148143 https://doi.org/10.1111/j.1467-937X.2005.00353.x

PRIMICERI, G. E. and TAMBALOTTI, A. (2020). Macroeconomic forecasting in the time of COVID-19. In *Manuscript*, *Northwestern University*.

REIMERS, H.-E. (1997). Forecasting of seasonal cointegrated processes. *Int. J. Forecast.* **13** 369–380.

ROSSI, B. and SEKHPOSYAN, T. (2014). Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. *Int. J. Forecast.* **30** 662–682. https://doi.org/10.1016/j.ijforecast.2013.03.005

SCHORFHEIDE, F. and SONG, D. (2021). Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic Technical Report National Bureau of Economic Research.

TU, Y. and YI, Y. (2017). Forecasting cointegrated nonstationary time series with time-varying variance. *J. Econometrics* **196** 83–98. MR3572815 https://doi.org/10.1016/j.jeconom.2016.09.012

WRÓBLEWSKA, J. (2023). Bayesian analysis of seasonally cointegrated VAR models. *Econom. Stat.*.

WRÓBLEWSKA, J. and PAJOR, A. (2019). One-period joint forecasts of Polish inflation, unemployment and interest rate using Bayesian VEC-MSF models. *Central Eur. J. Econ. Modell. Econom.* **11** 23–45.

WU, J. C. and XIA, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *J. Money Credit Bank.* **48** 253–291.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 235–260
https://doi.org/10.1214/24-AOAS1957

ESTIMATING PRODUCT CANNIBALISATION IN WHOLESALE USING MULTIVARIATE HAWKES PROCESSES WITH INHIBITION

BY ISABELLA DEUTSCH[1,a] AND GORDON J. ROSS[1,b]

[1]*School of Mathematics, University of Edinburgh,* [a]*deutsch.isabella@outlook.com,* [b]*gordon.ross@ed.ac.uk*

Product cannibalisation in the marketplace refers to the decrease in the sales of one product due to competition from another product. We examine this phenomenon in a wholesale data set provided by a major international company. We use a multivariate Hawkes process where each product is represented by a dimension, with cross-inhibition effects that model product cannibalisation. To implement the Hawkes process with inhibition, we resolve challenges regarding the integration of the intensity function and introduce a new, less restrictive condition for stability, as existing conditions are unnecessarily strict under inhibition. We conduct our analysis in a Bayesian framework for which we design a dimension-independent prior on the cross-inhibition based on a reparametrisation.

REFERENCES

AGUILAR-PALACIOS, C., MUÑOZ-ROMERO, S. and ROJO-ÁLVAREZ, J. L. (2021). Causal quantification of cannibalization during promotional sales in grocery retail. *IEEE Access* **9** 34078–34089.

APOSTOLOPOULOU, I., LINDERMAN, S., MILLER, K. and DUBRAWSKI, A. (2019). *Mutually Regressive Point Processes*. *Advances in Neural Information Processing Systems* **32**. Curran Associates, Red Hook.

ATASU, A., GUIDE, V. D. R. and VAN WASSENHOVE, L. N. (2010). So what if remanufacturing cannibalizes my new product sales? *Calif. Manag. Rev.* **52** 56–76.

BACRY, E., BOMPAIRE, M., GAÏFFAS, S. and MUZY, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *J. Mach. Learn. Res.* **21** Paper No. 50, 32. MR4095329

BACRY, E., JAISSON, T. and MUZY, J. (2016). Estimation of slowly decreasing Hawkes kernels: Application to high-frequency order book dynamics. *Quant. Finance* **16** 1179–1201. MR3520321 https://doi.org/10.1080/14697688.2015.1123287

BACRY, E., MASTROMATTEO, I. and MUZY, J.-F. (2015). Hawkes processes in finance. *Mark. Microstruct. Liq.* **1** 1550005.

BACRY, E. and MUZY, J.-F. (2016). First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Trans. Inf. Theory* **62** 2184–2202. MR3480107 https://doi.org/10.1109/TIT.2016.2533397

BEKAL, G. and BARI, M. (2021). An XGBoost-Based Forecasting Framework for Product Cannibalization. arXiv.

BLUNDELL, C., BECK, J. and HELLER, K. A. (2012). *Modelling Reciprocating Relationships with Hawkes Processes*. *Advances in Neural Information Processing Systems* **25**. Curran Associates, Red Hook.

BONNET, A., HERRERA, M. M. and SANGNIER, M. (2021). Maximum likelihood estimation for Hawkes processes with self-excitation or inhibition. *Statist. Probab. Lett.* **179** Paper No. 109214, 7. MR4304275 https://doi.org/10.1016/j.spl.2021.109214

BRÉMAUD, P. and MASSOULIÉ, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Probab.* **24** 1563–1588. MR1411506 https://doi.org/10.1214/aop/1065725193

BROWNING, R., SULEM, D., MENGERSEN, K., RIVOIRARD, V. and ROUSSEAU, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. *PLoS ONE* **16** e0250015. https://doi.org/10.1371/journal.pone.0250015

CHEN, F. and STINDL, T. (2018). Direct likelihood evaluation for the renewal Hawkes process. *J. Comput. Graph. Statist.* **27** 119–131. MR3788306 https://doi.org/10.1080/10618600.2017.1341324

CHILD, P., DIEDERICHS, R., SANDERS, F.-H. and WISNIOWSKI, S. (1991). SMR forum: The management of complexity. *Sloan Manage. Rev.* **33** 73–80.

COPULSKY, W. (1976). Cannibalism in the marketplace. *J. Mark.* **40** 103–105.

COSTA, M., GRAHAM, C., MARSALLE, L. and TRAN, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Adv. in Appl. Probab.* **52** 879–915. MR4153592 https://doi.org/10.1017/apr.2020.19

Key words and phrases. Intensity function, point process, prior choice, stability.

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

DE GIOVANNI, P. and RAMANI, V. (2018). Product cannibalization and the effect of a service strategy. *J. Oper. Res. Soc.* **69** 340–357.

DESAI, P. S. (2001). Quality segmentation in spatial markets: When does cannibalization affect product line design? *Mark. Sci.* **20** 265–283.

DEUTSCH, I. and ROSS, G. J. (2025). Supplement to "Estimating Product Cannibalisation in Wholesale using Multivariate Hawkes Processes with Inhibition." https://doi.org/10.1214/24-AOAS1957SUPPA, https://doi.org/10.1214/24-AOAS1957SUPPB

EICHLER, M. (2013). Causal inference with multiple time series: Principles and problems. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20110613, 17. MR3081173 https://doi.org/10.1098/rsta.2011.0613

EICHLER, M., DAHLHAUS, R. and DUECK, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *J. Time Series Anal.* **38** 225–242. MR3611742 https://doi.org/10.1111/jtsa.12213

EMBRECHTS, P., LINIGER, T. and LIN, L. (2011). Multivariate Hawkes processes: An application to financial data. *J. Appl. Probab.* **48A** 367–378. MR2865638 https://doi.org/10.1239/jap/1318940477

ERTEKIN, Ş., RUDIN, C. and MCCORMICK, T. H. (2015). Reactive point processes: A new approach to predicting power failures in underground electrical systems. *Ann. Appl. Stat.* **9** 122–144. MR3341110 https://doi.org/10.1214/14-AOAS789

GARNIER, R. (2022). Concurrent neural network: A model of competition between times series. *Ann. Oper. Res.* **313** 945–964. MR4440775 https://doi.org/10.1007/s10479-021-04253-3

GELFAND, I. (1941). Normierte Ringe. *Rec. Math.* [*Mat. Sb.*] *N.S.* **9** 3–24. MR0004726

GHOSE, A., SMITH, M. D. and TELANG, R. (2006). Internet exchanges for used books: An empirical analysis of product cannibalization and welfare impact. *Inf. Syst. Res.* **17** 3–19.

GUIDOLIN, M. and GUSEO, R. (2020). Has the iPhone cannibalized the iPad? An asymmetric competition model. *Appl. Stoch. Models Bus. Ind.* **36** 465–476. MR4111175 https://doi.org/10.1002/asmb.2505

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

HAWKES, A. G. and OAKES, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. MR0378093 https://doi.org/10.2307/3212693

HELMSTETTER, A., KAGAN, Y. Y. and JACKSON, D. D. (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *Bull. Seismol. Soc. Amer.* **96** 90–106.

HELMSTETTER, A. and SORNETTE, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J. Geophys. Res.*, *Solid Earth* **107** ESE 10-1–ESE 10-21.

KALAIR, K., CONNAUGHTON, C. and ALAIMO DI LORO, P. (2021). A non-parametric Hawkes process model of primary and secondary accidents on a UK smart motorway. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 80–97. MR4204939 https://doi.org/10.1111/rssc.12450

KAMAKURA, W. A. and SRIVASTAVA, R. K. (1984). Predicting choice shares under conditions of brand interdependence. *J. Mark. Res.* **21** 420–434.

KOLEV, A. A. and ROSS, G. J. (2019). Inference for ETAS models with non-Poissonian mainshock arrival times. *Stat. Comput.* **29** 915–931. MR3994609 https://doi.org/10.1007/s11222-018-9845-z

KONG, E. L. (2015). Cannibalization Effects of Products in Zara's Stores and Demand Forecasting. PhD thesis, Massachusetts Institute of Technology.

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

LEMONNIER, R. and VAYATIS, N. (2014). Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In *Proceedings of the* 2014*th European Conference on Machine Learning and Knowledge Discovery in Databases—Part II* 161–176. Springer, Berlin.

LU, X. and ABERGEL, F. (2018). High-dimensional Hawkes processes for limit order books: Modelling, empirical analysis and numerical calibration. *Quant. Finance* **18** 249–264. MR3750734 https://doi.org/10.1080/14697688.2017.1403142

MARKWICK, D. (2020). Bayesian Nonparametric Hawkes Processes with Applications. PhD thesis, Univ. College London.

MEI, H. and EISNER, J. (2017). *The Neural Hawkes Process*: *A Neurally Self-Modulating Multivariate Point Process*. *Advances in Neural Information Processing Systems* **30**. Curran Associates, Red Hook.

MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. MR3127957 https://doi.org/10.1214/13-AOAS647

MOLKENTHIN, C., DONNER, C., REICH, S., ZÖLLER, G., HAINZL, S., HOLSCHNEIDER, M. and OPPER, M. (2022). GP-ETAS: Semiparametric Bayesian inference for the spatio-temporal epidemic type aftershock sequence model. *Stat. Comput.* **32** Paper No. 29, 25. MR4394859 https://doi.org/10.1007/s11222-022-10085-3

OGATA, Y. (1981). On Lewis' simulation method for point processes. *IEEE Trans. Inf. Theory* **27** 23–31.

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

OKORIE, O., OBI, M., RUSSELL, J., CHARNLEY, F. and SALONITIS, K. (2021). A triple bottom line examination of product cannibalisation and remanufacturing: A review and research agenda. *Sustain. Prod. Consump.* **27** 958–974.

PITKIN, J., MANOLOPOULOU, I. and ROSS, G. (2024). Bayesian hierarchical modelling of sparse count processes in retail analytics. *Ann. Appl. Stat.* **18** 946–965. MR4728650 https://doi.org/10.1214/23-aoas1811

RAMBALDI, M., BACRY, E. and LILLO, F. (2017). The role of volume in order book dynamics: A multivariate Hawkes process analysis. *Quant. Finance* **17** 999–1020. MR3660377 https://doi.org/10.1080/14697688.2016.1260759

RASMUSSEN, J. G. (2013). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642. MR3085883 https://doi.org/10.1007/s11009-011-9272-5

ROSS, G. J. (2021). Bayesian estimation of the ETAS model for earthquake occurrences. *Bull. Seismol. Soc. Amer.* **111** 1473–1480.

ROSS, G. J. and KOLEV, A. A. (2022). Semiparametric Bayesian forecasting of spatiotemporal earthquake occurrences. *Ann. Appl. Stat.* **16** 2083–2100. MR4489200 https://doi.org/10.1214/21-aoas1554

RUIZ, F. J. R., ATHEY, S. and BLEI, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *Ann. Appl. Stat.* **14** 1–27. MR4085081 https://doi.org/10.1214/19-AOAS1265

SCHOENBERG, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *J. Amer. Statist. Assoc.* **98** 789–795. MR2055487 https://doi.org/10.1198/016214503000000710

SERRE, D. (2002). *Matrices: Theory and Applications. Graduate Texts in Mathematics* **216**. Springer, New York. Translated from the 2001 French original. MR1923507

SHELTON, C., QIN, Z. and SHETTY, C. (2018). Hawkes process inference with missing data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* AAAI Press, Menlo Park.

SULEM, D., RIVOIRARD, V. and ROUSSEAU, J. (2024). Bayesian estimation of nonlinear Hawkes processes. *Bernoulli* **30** 1257–1286. MR4699552 https://doi.org/10.3150/23-bej1631

TUCKER, J. D., SHAND, L. and LEWIS, J. R. (2019). Handling missing data in self-exciting point process models. *Spat. Stat.* **29** 160–176. MR3895111 https://doi.org/10.1016/j.spasta.2018.12.004

US CENSUS BUREAU (2011). Section 22. Wholesale and Retail Trade. Statistical Abstract of the United States. Government Printing Office.

VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998 https://doi.org/10.1198/016214508000000148

STAN DEVELOPMENT TEAM (2019). RStan: the R interface to Stan. R package version 2.19.2.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 261–285
https://doi.org/10.1214/24-AOAS1958

POISSON CLUSTER PROCESS MODELS FOR DETECTING ULTRA-DIFFUSE GALAXIES

BY DAYI LI[1,a], ALEX STRINGER[2,d], PATRICK E. BROWN[1,b], GWENDOLYN M. EADIE[1,3,c] AND ROBERTO G. ABRAHAM[3,e]

[1]*Department of Statistical Sciences, University of Toronto,* [a]*dayi.li@mail.utoronto.ca,* [b]*patrick.brown@utoronto.ca,* [c]*gwen.eadie@utoronto.ca*

[2]*Department of Statistics and Actuarial Science, University of Waterloo,* [d]*alex.stringer@uwaterloo.ca*

[3]*David A. Dunlap Department of Astronomy & Astrophysics, University of Toronto,* [e]*roberto.abraham@utoronto.ca*

We propose a novel set of Poisson cluster process (PCP) models to detect ultra-diffuse galaxies (UDGs), a class of extremely faint, enigmatic galaxies of substantial interest in modern astrophysics. We model the unobserved UDG locations as parent points in a PCP and infer their positions based on the observed spatial point patterns of their old star cluster systems. Many UDGs have somewhere from a few to hundreds of these old star clusters, which we treat as offspring points in our models. We also present a new framework to construct a marked PCP model using the marks/characteristics of star clusters (their colors, brightnesses, etc.). The marked PCP model may enhance the detection of UDGs and offers broad applicability to problems in other disciplines. To assess the overall model performance, we design an innovative assessment tool for spatial prediction problems where only point-referenced ground truth is available, overcoming the limitation of standard ROC analyses where spatial Boolean reference maps are required. We construct a bespoke blocked Gibbs adaptive spatial birth-death-move Markov chain Monte Carlo algorithm to infer the locations of UDGs using real data from a *Hubble Space Telescope* imaging survey. Based on our performance assessment tool, our novel models significantly outperform existing approaches using the log-Gaussian Cox process. We also obtained preliminary evidence that the marked PCP model may improve UDG detection performance compared to the model without marks. Furthermore, we find evidence of a potential new "dark galaxy" that was not detected by previous methods.

REFERENCES

AMORISCO, N. C., MONACHESI, A., AGNELLO, A. and WHITE, S. D. M. (2018). The globular cluster systems of 54 coma ultra-diffuse galaxies: Statistical constraints from hst data. *Mon. Not. R. Astron. Soc.* **475** 4235–4251.

BADDELEY, A., DAVIES, T. M., HAZELTON, M. L., RAKSHIT, S. and TURNER, R. (2022). Fundamental problems in fitting spatial cluster process models. *Spat. Stat.* **52** Paper No. 100709. MR4499282

BERTIN, E. and ARNOUTS, S. (1996). SExtractor: Software for source extraction. *Astron. Astrophys. Suppl. Ser.* **117** 393–404.

BILODEAU, B., STRINGER, A. and TANG, Y. (2024). Stochastic convergence rates and applications of adaptive quadrature in Bayesian inference. *J. Amer. Statist. Assoc.* **119** 690–700. MR4713923

BOLIN, D. and LINDGREN, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 85–106. MR3299400

BOLIN, D. and LINDGREN, F. (2018). Calculating probabilistic excursion sets and related quantities using excursions. *J. Stat. Softw.* **86** 1–20.

BU, F., KAGAAYI, J., GRABOWSKI, M. K., RATMANN, O. and XU, J. (2024). Inferring HIV transmission patterns from viral deep-sequence data via latent typed point processes. *Biometrics* **80**.

BURKERT, A. (2017). The geometry and origin of ultra-diffuse ghost galaxies. *Astrophys. J.* **838** 93.

BURKERT, A. (2020). Fuzzy dark matter and dark matter halo cores. *Astrophys. J.* **904** 161.

Key words and phrases. Point process, Poisson cluster process, cluster detection, ultra-diffuse galaxies, globular clusters, astrostatistics.

CASTELLOE, J. M. (1998). Issues in reversible jump Markov chain Monte Carlo and composite EM analysis, applied to spatial Poisson cluster processes. Ph.D. thesis, University of Iowa.

CHOWDHURY, D. D., VAN DEN BOSCH, F. C. and VAN DOKKUM, P. (2019). On the orbital decay of globular clusters in NGC 1052-DF2: Testing a baryon-only mass model. *Astrophys. J.* **877** 133.

DANIELI, S., VAN DOKKUM, P., TRUJILLO-GOMEZ, S., KRUIJSSEN, J. M. D., ROMANOWSKY, A. J., CARLSTEN, S., SHEN, Z., LI, J., ABRAHAM, R. et al. (2022). Ngc 5846-udg1: A galaxy formed mostly by star formation in massive, extremely dense clumps of gas. *Astrophys. J. Lett.* **927** L28.

DIGGLE, P. J., MENEZES, R. and SU, T. (2010). Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 191–232. MR2744471

DIGGLE, P. J., MORAGA, P., ROWLINGSON, B. and TAYLOR, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statist. Sci.* **28** 542–563. MR3161587

DYTSO, A., BUSTIN, R., POOR, H. V. and SHAMAI, S. (2018). Analytical properties of generalized Gaussian distributions. *J. Stat. Distrib. Appl.* **5** 6.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27** 861–874.

FORBES, D. A. (2017). How large are the globular cluster systems of early-type galaxies and do they scale with galaxy halo properties? *Mon. Not. R. Astron. Soc.* **472** L104–L108.

FORBES, D. A., ALABI, A., ROMANOWSKY, A. J., BRODIE, J. P. and ARIMOTO, N. (2020). Globular clusters in Coma cluster ultra-diffuse galaxies (UDGs): Evidence for two types of UDG? *Mon. Not. R. Astron. Soc.* **492** 4874–4883.

FORBES, D. A., BASTIAN, N., GIELES, M. et al. (2018). Globular cluster formation and evolution in the context of cosmological galaxy assembly: Open questions. *Proc. R. Soc. Edinb. A* **474** 20170616. MR3782772

GEYER, C. J. and MØLLER, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Stat.* **21** 359–373. MR1310082

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GUDEHUS, D. (1995). The virgocentric velocity from bright cluster galaxies. *Astron. Astrophys.* **302** 21.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HARRIS, W. (1991). Globular cluster systems in galaxies beyond the local group. *Annu. Rev. Astron. Astrophys.* **29** 543–579.

HARRIS, W. E., BROWN, R. A., DURRELL, P. R., ROMANOWSKY, A. J., BLAKESLEE, J., BRODIE, J., JANSSENS, S., LISKER, T., OKAMOTO, S. et al. (2020). The PIPER survey. I. An initial look at the intergalactic globular cluster population in the perseus cluster. *Astrophys. J.* **890** 105.

HARRIS, W. E. and RACINE, R. (1979). Globular clusters in galaxies. *Annu. Rev. Astron. Astrophys.* **17** 241–274.

HO, L. P. and STOYAN, D. (2008). Modelling marked point patterns by intensity-marked Cox processes. *Statist. Probab. Lett.* **78** 1194–1199. MR2441462

HONG, C. (2022). Deep Neyman-Scott Processes. Ph.D. thesis, University of California, Riverside.

HONG, C. and SHELTON, C. R. (2023). Variational inference for Neyman-Scott processes.

HUI, L., OSTRIKER, J. P., TREMAINE, S. and WITTEN, E. (2017). Ultralight scalars as cosmological dark matter. *Phys. Rev. D* **95** 043541.

JANSSENS, S. R., ABRAHAM, R., BRODIE, J., FORBES, D. A. and ROMANOWSKY, A. J. (2019). The distribution of ultra-diffuse and ultra-compact galaxies in the Frontier fields. *Astrophys. J.* **887** 92.

JANSSENS, S. R., ROMANOWSKY, A. J., ABRAHAM, R., BRODIE, J. P., COUCH, W. J., FORBES, D. A., LAINE, S., MARTÍNEZ-DELGADO, D. and VAN DOKKUM, P. G. (2022). The globular clusters and star formation history of the isolated, quiescent ultra-diffuse galaxy dgsat-i. *Mon. Not. R. Astron. Soc.* **517** 858–871.

LI, D., STRINGER, A., BROWN, P. E., EADIE, G. M. and ABRAHAM, R. G. (2025). Supplement to "Poisson Cluster Process Models for Detecting Ultra-Diffuse Galaxies." https://doi.org/10.1214/24-AOAS1958SUPPA, https://doi.org/10.1214/24-AOAS1958SUPPB, https://doi.org/10.1214/24-AOAS1958SUPPC

LI, D. D., EADIE, G. M., ABRAHAM, R., BROWN, P. E., HARRIS, W. E., JANSSENS, S. R., ROMANOWSKY, A. J., VAN DOKKUM, P. and DANIELI, S. (2022). Light from the darkness: Detecting ultra-diffuse galaxies in the perseus cluster through over-densities of globular clusters with a log-Gaussian Cox process. *Astrophys. J.* **935** 3.

LIM, S., CÔTÉ, P., PENG, E. W., FERRARESE, L., ROEDIGER, J. C., DURRELL, P. R., MIHOS, J. C., WANG, K., GWYN, S. D. J. et al. (2020). The next generation virgo cluster survey (NGVS). XXX. Ultra-diffuse galaxies and their globular cluster systems. *Astrophys. J.* **899** 69.

LIM, S., PENG, E. W., CÔTÉ, P., SALES, L. V., DEN BROK, M., BLAKESLEE, J. P. and GUHATHAKURTA, P. (2018). The globular cluster systems of ultra-diffuse galaxies in the Coma cluster. *Astrophys. J.* **862** 82.

LOTZ, J. M., MILLER, B. W. and FERGUSON, H. C. (2004). The colors of dwarf elliptical galaxy globular cluster systems, nuclei, and stellar halos. *Astrophys. J.* **613** 262–278.

MADAU, P., LUPI, A., DIEMAND, J., BURKERT, A. and LIN, D. N. C. (2020). Globular cluster formation from colliding substructure. *Astrophys. J.* **890** 18.

MØLLER, J. (2003). Shot noise Cox processes. *Adv. in Appl. Probab.* **35** 614–640. MR1990607 https://doi.org/10.1239/aap/1059486821

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019

MØLLER, J. and TOFTAKER, H. (2014). Geometric anisotropic spatial point pattern analysis and Cox processes. *Scand. J. Stat.* **41** 414–435. MR3207179

MØLLER, J. and TORRISI, G. L. (2005). Generalised shot noise Cox processes. *Adv. in Appl. Probab.* **37** 48–74. MR2135153

MYLLYMÄKI, M. and PENTTINEN, A. (2009). Conditionally heteroscedastic intensity-dependent marking of log Gaussian Cox processes. *Stat. Neerl.* **63** 450–473. MR2598981

NEYMAN, J. and SCOTT, E. L. (1958). Statistical approach to problems of cosmology. *J. Roy. Statist. Soc. Ser. B* **20** 1–43. MR0105309

PENG, E. W., JORDÁN, A., CÔTÉ, P., BLAKESLEE, J. P., FERRARESE, L., MEI, S., WEST, M. J., MERRITT, D., MILOSAVLJEVIĆ, M. et al. (2006). The ACS virgo cluster survey. IX. The color distributions of globular cluster systems in early-type galaxies*. *Astrophys. J.* **639** 95.

PENG, E. W. and LIM, S. (2016). A rich globular cluster system in Dragonfly-17: Are ultra-diffuse galaxies pure stellar halos?*. *Astrophys. J. Lett.* **822** L31.

REJKUBA, M. (2012). Globular cluster luminosity function as distance indicator. *Astrophys. Space Sci.* **341** 195–206.

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16** 351–367. MR1888450

ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. MR2749836

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

SAIFOLLAHI, T., ZARITSKY, D., TRUJILLO, I., PELETIER, R. F., KNAPEN, J. H., AMORISCO, N., BEASLEY, M. A. and DONNERSTEIN, R. (2022). Implications for galaxy formation models from observations of globular clusters around ultradiffuse galaxies. *Mon. Not. R. Astron. Soc.* **511** 4633–4659.

SÉRSIC, J. L. (1963). Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *BAAA* **6** 41–43.

SHEN, Z., DANIELI, S., VAN DOKKUM, P., ABRAHAM, R., BRODIE, J. P., CONROY, C., DOLPHIN, A. E., ROMANOWSKY, A. J., KRUIJSSEN, J. M. D. et al. (2021). A tip of the red giant branch distance of 22.1 ± 1.2 mpc to the dark matter deficient galaxy NGC 1052–DF2 from 40 orbits of Hubble space telescope imaging. *Astrophys. J. Lett.* **914** L12.

SHEN, Z., VAN DOKKUM, P. and DANIELI, S. (2021). A complex luminosity function for the anomalous globular clusters in NGC 1052-DF2 and NGC 1052-DF4. *Astrophys. J.* **909** 179.

TANG, Y., ROMANOWSKY, A. J., VAN DOKKUM, P. G., JARRETT, T. H. and JANSSENS, S. R. (2024). Observational constraints on the formation of dark-matter deficient galaxies from morphologies, stellar populations and globular clusters. In *Proceedings of DGSCS*.

VAN DOKKUM, P., ABRAHAM, R., BRODIE, J., CONROY, C., DANIELI, S., MERRITT, A., MOWLA, L., ROMANOWSKY, A. and ZHANG, J. (2016). A high stellar velocity dispersion and $\sim$100 globular clusters for the ultra-diffuse galaxy Dragonfly 44. *Astrophys. J. Lett.* **828** L6.

VAN DOKKUM, P., ABRAHAM, R., ROMANOWSKY, A. J., BRODIE, J., CONROY, C., DANIELI, S., LOKHORST, D., MERRITT, A., MOWLA, L. et al. (2017). Extensive globular cluster systems associated with ultra diffuse galaxies in the coma cluster. *Astrophys. J. Lett.* **844** L11.

VAN DOKKUM, P., SHEN, Z., ROMANOWSKY, A. J., ABRAHAM, R., CONROY, C., DANIELI, S., CHOWDHURY, D. D., KEIM, M. A., KRUIJSSEN, J. M. D. et al. (2022). Monochromatic globular clusters as a critical test of formation models for the dark matter–deficient galaxies NGC 1052-DF2 and NGC 1052-DF4. *Astrophys. J. Lett.* **940** L9.

VAN DOKKUM, P., WASSERMAN, A., DANIELI, S., ABRAHAM, R., BRODIE, J., CONROY, C., FORBES, D. A., MARTIN, C., MATUSZEWSKI, M. et al. (2019). Spatially resolved stellar kinematics of the ultra-diffuse galaxy Dragonfly 44. I. Observations, kinematics, and cold dark matter halo fits. *Astrophys. J.* **880** 91.

VAN DOKKUM, P. G., ABRAHAM, R., MERRITT, A., ZHANG, J., GEHA, M. and CONROY, C. (2015). Forty-seven Milky Way-sized, extremely diffuse galaxies in the Coma cluster. *Astrophys. J. Lett.* **798** L45.

WAAGEPETERSEN, R., GUAN, Y., JALILIAN, A. and MATEU, J. (2016). Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 77–96. MR3438239

WAAGEPETERSEN, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63** 252–258. MR2345595

WANG, Q., PENG, E. W., BLAKESLEE, J. P., CÔTÉ, P., FERRARESE, L., JORDÁN, A., MEI, S. and WEST, M. J. (2013). The acs virgo cluster survey. XVII. The spatial alignment of globular cluster systems with early-type host galaxies. *Astrophys*. *J*. **769** 145.

WANG, Y., DEGLERIS, A., WILLIAMS, A. and LINDERMAN, S. W. (2024). Spatiotemporal clustering with Neyman-Scott processes via connections to Bayesian nonparametric mixture models. *J. Amer. Statist*. *Assoc*. **119** 2382–2395. MR4797948

WASSERMAN, A., VAN DOKKUM, P., ROMANOWSKY, A. J., BRODIE, J., DANIELI, S., FORBES, D. A., ABRAHAM, R., MARTIN, C., MATUSZEWSKI, M. et al. (2019). Spatially resolved stellar kinematics of the ultra-diffuse galaxy Dragonfly 44. II. Constraints on fuzzy dark matter. *Astrophys*. *J*. **885** 155.

WITTMANN, C., LISKER, T., TILAHUN, L. A., GREBEL, E. K., CONSELICE, C. J., PENNY, S., JANZ, J., GALLAGHER, J. S., KOTULLA, R. et al. (2017). A population of faint low surface brightness galaxies in the Perseus cluster core. *Mon*. *Not*. *R*. *Astron*. *Soc*. **470** 1512–1525.

YAGI, M., KODA, J., KOMIYAMA, Y. and YAMANOI, H. (2016). Catalog of ultra-diffuse galaxies in the Coma clusters from Subaru imaging data*. *Astrophys*. *J*., *Suppl*. *Ser*. **225** 11.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 286–301
https://doi.org/10.1214/24-AOAS1959

THE SHORT-TERM DYNAMICS OF CONFLICT-DRIVEN DISPLACEMENT: BAYESIAN MODELING OF DISAGGREGATED DATA FROM SOMALIA

BY GREGOR ZENS[1,a] AND LISA THALHEIMER[2,b]

[1]*Population and Just Societies Program, International Institute for Applied Systems Analysis (IIASA),* [a]*zens@iiasa.ac.at*
[2]*Environment and Migration: Interactions and Choices Division, United Nations University (UNU),* [b]*thalheimer@ehs.unu.edu*

Understanding the short-run dynamics of conflict and forced displacement is crucial for the design of effective policy responses, yet quantitative analyses in this realm are sparse. This is primarily due to the scarcity of high-frequency displacement data and methodological challenges arising when modeling imperfect data collected in conflict zones. Addressing both issues, we develop a Bayesian panel regression model to assess the short-term impact of conflict on displacement in Somalia, utilizing weekly panel data that encompasses eight million displacements and 19,000 conflict events from 2017 to 2023. Results suggest a rapid and nonlinear displacement response postconflict, with significant heterogeneity in effects dependent on the nature of conflict events. In a displacement forecasting exercise, our model outperforms standard benchmarks, underscoring its potential for informing decision-makers in crisis scenarios.

REFERENCES

ABEL, G. J., BROTTRAGER, M., CUARESMA, J. C. and MUTTARAK, R. (2019). Climate, conflict and forced migration. *Glob. Environ. Change* **54** 239–249.

BALL, P. and ASHER, J. (2002). Statistics and Slobodan: Using data analysis and statistics in the war crimes trial of former president Milosevic. *Chance* **15** 17–24. MR1930510 https://doi.org/10.1080/09332480.2002.10554820

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

BEYER, R. M., SCHEWE, J. and LOTZE-CAMPEN, H. (2022). Gravity models do not explain, and cannot predict, international migration dynamics. *Humanit. Soc. Sci. Commun.* **9** 1–10.

BIJAK, J., DISNEY, G., FINDLAY, A. M., FORSTER, J. J., SMITH, P. W. F. and WIŚNIOWSKI, A. (2019). Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *J. Forecast.* **38** 470–487. MR4002373 https://doi.org/10.1002/for.2576

BOHRA-MISHRA, P. and MASSEY, D. S. (2011). Individual decisions to migrate during civil conflict. *Demography* **48** 401–424. https://doi.org/10.1007/s13524-011-0016-5

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

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80. PMLR.

CHAN, J. C. C. and JELIAZKOV, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *Int. J. Math. Model. Numer. Optim.* **1** 101–120.

CONTE, A. and MIGALI, S. (2019). The role of conflict and organized violence in international forced migration. *Demogr. Res.* **41** 393–424.

CONTI, G., FRÜHWIRTH-SCHNATTER, S., HECKMAN, J. J. and PIATEK, R. (2014). Bayesian exploratory factor analysis. *J. Econometrics* **183** 31–57. MR3269916 https://doi.org/10.1016/j.jeconom.2014.06.008

DAVENPORT, C., MOORE, W. and POE, S. (2003). Sometimes you just have to leave: Domestic threats and forced migration, 1964-1989. *Int. Interact.* **29** 27–55.

Key words and phrases. Internal displacement modeling, factor model, state space model, distributed lag model, panel model.

DIEBOLD, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. *J. Bus. Econom. Statist.* **33** 1–9. MR3303732 https://doi.org/10.1080/07350015.2014.983236

DIRIKGIL, N. (2023). Addressing the prevention of internal displacement: The right not to be arbitrarily displaced. *J. Int. Migr. Integr.* **24** 113–138.

DRC (2022). Global displacement forecast 2022. Danish Refugee Council Technical Report.

ECK, K. (2012). In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Coop. Confl.* **47** 124–141.

ENGEL, S. and IBÁÑEZ, A. M. (2007). Displacement due to violence in Colombia: A household-level analysis. *Econ. Dev. Cult. Change* **55** 335–365.

FRÜHWIRTH-SCHNATTER, S., HOSSZEJNI, D. and LOPES, H. F. (2024). Sparse Bayesian factor analysis when the number of factors is unknown. *Bayesian Anal.* **1** 1–31.

GHYSELS, E., SINKO, A. and VALKANOV, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Rev.* **26** 53–90. MR2339264 https://doi.org/10.1080/07474930600972467

HATTON, T. J. (2009). The rise and fall of asylum: What happened and why? *Econ. J.* **119** F183–F213.

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

LU, X., BENGTSSON, L. and HOLME, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Natl. Acad. Sci. USA* **109** 11576–11581.

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

MARTIN, S. F. and SINGH, L. (2019). Big data and early warning of displacement. In *Mobilizing Global Knowledge: Refugee Research in an Age of Displacement* 129–150.

MCCAUSLAND, W. J., MILLER, S. and PELLETIER, D. (2011). Simulation smoothing for state-space models: A computational efficiency analysis. *Comput. Statist. Data Anal.* **55** 199–212. MR2736547 https://doi.org/10.1016/j.csda.2010.07.009

OH, W. S., MUNEEPEERAKUL, R., RUBENSTEIN, D. and LEVIN, S. (2024). Emergent network patterns of internal displacement in Somalia driven by natural disasters and conflicts. *Glob. Environ. Change* **84** 102793.

PHAM, K. H. and LUENGO-OROZ, M. (2023). Predictive modeling of movements of refugees and internally displaced people: Towards a computational framework. *J. Ethn. Migr. Stud.* **49** 408–444.

PIIRONEN, J. and VEHTARI, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11** 5018–5051. MR3738204 https://doi.org/10.1214/17-EJS1337SI

RALEIGH, C., LINKE, A., HEGRE, H. and KARLSEN, J. (2010). Introducing ACLED: An armed conflict location and event dataset: Special data feature. *J. Peace Res.* **47** 651–660.

SARZIN, Z. I. (2017). Stocktaking of global forced displacement data. World Bank Policy Research Working Paper **7985**.

SCHMEIDL, S. (1997). Exploring the causes of forced migration: A pooled time-series analysis, 1971–1990. *Soc. Sci. Q.* **78** 284–308.

SCHMEIDL, S. (2001). Conflict and forced migration: A quantitative review, 1964–1995. In *Global Migrants, Global Refugees: Problems and Solutions* 62–94.

SCHON, J. (2015). Focus on the forest, not the trees: A changepoint model of forced displacement. *J. Refug. Stud.* **28** 437–467.

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

STEEL, M. F. and ZENS, G. (2024). Model uncertainty in latent Gaussian models with univariate link function. Preprint. Available at arXiv:2406.17318.

TAI, X. H., MEHRA, S. and BLUMENSTOCK, J. E. (2022). Mobile phone data reveal the effects of violence on internal displacement in Afghanistan. *Nat. Hum. Behav.* **6** 624–634. https://doi.org/10.1038/s41562-022-01336-4

THALHEIMER, L., SCHWARZ, M. P. and PRETIS, F. (2023). Large weather and conflict effects on internal displacement in Somalia with little evidence of feedback onto conflict. *Glob. Environ. Change* **79** 102641.

THALHEIMER, L., SIMPERINGHAM, E. and JJEMBA, E. W. (2022). The role of anticipatory humanitarian action to reduce disaster displacement. *Environ. Res. Lett.* **17** 014043.

UNHCR (2017). Internal displacements recorded by Protection and Return Monitoring Network Somalia. Notes on methodology.

WAGNER, H., FRÜHWIRTH-SCHNATTER, S. and JACOBI, L. (2023). Factor-augmented Bayesian treatment effects models for panel outcomes. *Econom. Stat.* **28** 63–80. MR4644292 https://doi.org/10.1016/j.ecosta.2022.04.003

WELCH, N. G. and RAFTERY, A. E. (2022). Probabilistic forecasts of international bilateral migration flows. *Proc. Natl. Acad. Sci. USA* **119** e2203822119.

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

WYCOFF, N., ARAB, A., DONATO, K., SINGH, L., KAWINTIRANON, K., LIU, Y. and JACOBS, E. (2023). Forecasting Ukrainian refugee flows with organic data sources. *Int. Migr. Rev.* 01979183231203931.

ZENS, G. and THALHEIMER, L. (2025). Supplement to "The short-term dynamics of conflict-driven displacement: Bayesian modeling of disaggregated data from Somalia." https://doi.org/10.1214/24-AOAS1959SUPP

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 302–328
https://doi.org/10.1214/24-AOAS1960

ACCOUNTING FOR SHARED COVARIATES IN SEMIPARAMETRIC BAYESIAN ADDITIVE REGRESSION TREES

BY ESTEVÃO B. PRADO[1,a], ANDREW C. PARNELL[2,3,4,b], RAFAEL A. MORAL[2,3,c], NATHAN MCJAMES[2,3,d], ANN O'SHEA[3,f] AND KEEFE MURPHY[2,3,e]

[1]*School of Mathematical Sciences, Lancaster University,* [a]*e.prado1@lancaster.ac.uk*

[2]*Hamilton Institute, Maynooth University,* [b]*andrew.parnell@mu.ie,* [c]*rafael.deandrademoral@mu.ie,* [d]*nathan.mcjames.2016@mumail.ie,* [e]*keefe.murphy@mu.ie*

[3]*Department of Mathematics & Statistics, Maynooth University,* [f]*ann.oshea@mu.ie*

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

We propose some extensions to semiparametric models based on Bayesian additive regression trees (BART). In the semiparametric BART paradigm, the response variable is approximated by a linear predictor and a BART model, where the linear component is responsible for estimating the main effects and BART accounts for nonspecified interactions and nonlinearities. Previous semiparametric models based on BART have assumed that the set of covariates in the linear predictor and the BART model are mutually exclusive in an attempt to avoid poor coverage properties and reduce bias in the estimates of the parameters in the linear predictor. The main novelty in our approach lies in the way we change the tree-generation moves in BART to deal with this bias and resolve nonidentifiability issues between the parametric and nonparametric components, even when they have covariates in common. This allows us to model complex interactions involving the covariates of primary interest, both among themselves and with those in the BART component. Our novel method is developed with a view to analysing data from an international education assessment, where certain predictors of students' achievements in mathematics are of particular interpretational interest. Through additional simulation studies and another application to a well-known benchmark dataset, we also show competitive performance when compared to regression models, alternative formulations of semiparametric BART, and other tree-based methods. The implementation of the proposed method is available at https://github.com/ebprado/CSP-BART.

REFERENCES

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

BEATON, A. E. et al. (1996). *Mathematics Achievement in the Middle School Years. IEA's Third International Mathematics and Science Study* (*TIMSS*). ERIC.

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

CHEN, X. (2022). The effects of individual-and class-level achievement on attitudes towards mathematics: An analysis of Hong Kong students using TIMSS 2019. *Stud. Educ. Eval.* **72** 101113.

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

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2013). Bayesian regression structure discovery. In *Bayesian Theory and Applications* (P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens, eds.) 451–465 22. Oxford Univ. Press, Oxford.

CHIPMAN, H. A., GEORGE, E. I., MCCULLOCH, R. E. and SHIVELY, T. S. (2022). mBART: Multidimensional monotone BART. *Bayesian Anal.* **17** 515–544. MR4483229 https://doi.org/10.1214/21-ba1259

DESHPANDE, S. K., BAI, R., BALOCCHI, C., STARLING, J. E. and WEISS, J. (2020). VCBART: Bayesian trees for varying coefficients. arXiv preprint. Available at arXiv:2003.06416.

Key words and phrases. Bayesian additive regression trees, Bayesian nonparametric regression, generalised linear models, semiparametric regression, Trends in International Mathematics and Science Study, TIMSS 2019.

DORIE, V. (2020). dbarts: Discrete Bayesian additive regression trees sampler. R package version 0.9-19.

DORIE, V., PERRETT, G., HILL, J. L. and GOODRICH, B. (2022). Stan and BART for causal inference: Estimating heterogeneous treatment effects using the power of Stan and the flexibility of machine learning. *Entropy* **24** 1782. MR4536778 https://doi.org/10.3390/e24121782

FISHBEIN, B., FOY, P. and YIN, L. (2021). TIMSS 2019 user guide for the international database Technical Report TIMSS & PIRLS International Study Center Boston, USA. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-database/.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67.

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

GOLDSTEIN, A., KAPELNER, A., BLEICH, J. and PITKIN, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Statist.* **24** 44–65. MR3328247 https://doi.org/10.1080/10618600.2014.907095

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.

GRILLI, L., PENNONI, F., RAMPICHINI, C. and ROMEO, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. *Ann. Appl. Stat.* **10** 2405–2426. MR3592062 https://doi.org/10.1214/16-AOAS988

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

HAREZLAK, J., RUPPERT, D. and WAND, M. P. (2018). *Semiparametric Regression with R. Use R!* Springer, New York. MR3889001 https://doi.org/10.1007/978-1-4939-8853-2

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HERNÁNDEZ, B., PENNINGTON, S. R. and PARNELL, A. C. (2015). Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics* **9** 54–64.

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

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

INGLIS, A., PARNELL, A. and HURLEY, C. B. (2022). Visualizing variable importance and variable interaction effects in machine learning models. *J. Comput. Graph. Statist.* **31** 766–778. MR4495709 https://doi.org/10.1080/10618600.2021.2007935

KAPELNER, A. and BLEICH, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *J. Stat. Softw.* **70** 1–40.

KINDO, B. P., WANG, H. and PEÑA, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat* **5** 119–131. MR3484083 https://doi.org/10.1002/sta4.110

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

LINERO, A. R., BASAK, P., LI, Y. and SINHA, D. (2022). Bayesian survival tree ensembles with submodel shrinkage. *Bayesian Anal.* **17** 997–1020. MR4505386 https://doi.org/10.1214/21-ba1285

LINERO, A. R., SINHA, D. and LIPSITZ, S. R. (2020). Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics* **76** 131–144. MR4098550 https://doi.org/10.1111/biom.13107

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

LOH, W.-Y. (2009). Improving the precision of classification trees. *Ann. Appl. Stat.* **3** 1710–1737. MR2752155 https://doi.org/10.1214/09-AOAS260

MARTIN, M. O., MULLIS, I. V., GREGORY, K. D., HOYLE, C. and SHEN, C. (2000). *Effective Schools in Science and Mathematics. IEA's Third International Mathematics and Science Study*. IEA, Chestnut Hill, MA.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series*. CRC Press, Boca Raton.

MCJAMES, N., PARNELL, A., GOH, Y. C. and O'SHEA, A. (2024). Bayesian causal forests for multivariate outcomes: Application to Irish data from an international large scale education assessment. *J. R. Stat. Soc., A.* **00** 1–23.

MOHAMMADPOUR, E., SHEKARCHIZADEH, A. and KALANTARRASHIDI, S. A. (2015). Multilevel modeling of science achievement in the TIMSS participating countries. *J. Educ. Res.* **108** 449–464.

MULLIS, I., MARTIN, M., FOY, P., KELLY, D. and FISHBEIN, B. (2020). TIMSS 2019 international results in mathematics and science Technical Report TIMSS & PIRLS International Study Center Boston, USA. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/.

MURRAY, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *J. Amer. Statist. Assoc.* **116** 756–769. MR4270022 https://doi.org/10.1080/01621459.2020.1813587

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *J. R. Stat. Soc., A* **135** 370–384.

NEWMAN, D. J., HETTICH, S., BLAKE, C. L. and MERZ, C. J. (1998). UCI repository of machine learning databases.

PRADO, E. B., MORAL, R. A. and PARNELL, A. C. (2021). Bayesian additive regression trees with model trees. *Stat. Comput.* **31** 20. MR4224945 https://doi.org/10.1007/s11222-021-09997-3

PRADO, E. B., PARNELL, A. C., MORAL, R. A., MCJAMES, N., O'SHEA, A. and MURPHY, K. (2025). Supplement to "Accounting for shared covariates in semiparametric Bayesian additive regression trees." https://doi.org/10.1214/24-AOAS1960SUPP

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

ROČKOVÁ, V. and SAHA, E. (2019). On theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics* **89** 2839–2848. PMLR.

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

SARTI, D. A., PRADO, E. B., INGLIS, A. N., DOS SANTOS, A. A. L., HURLEY, C. B., MORAL, R. A. and PARNELL, A. C. (2023). Bayesian additive regression trees for genotype by environment interaction models. *Ann. Appl. Stat.* **17** 1936–1957. MR4637651 https://doi.org/10.1214/22-aoas1698

SPARAPANI, R., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2020). Nonparametric competing risks analysis using Bayesian additive regression trees. *Stat. Methods Med. Res.* **29** 57–77. MR4055122 https://doi.org/10.1177/0962280218822140

SPARAPANI, R., SPANBAUER, C. and MCCULLOCH, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package.

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

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

STIGLER, J. W., GALLIMORE, R. and HIEBERT, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educ. Psychol.* **35** 87–100.

STIGLER, J. W. and HIEBERT, J. (1997). Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. *Phi Delta Kappan* **79** 14.

SUK, Y., KIM, J.-S. and KANG, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *J. Educ. Behav. Stat.* **46** 323–347.

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

TANG, A., LI, W. and LIU, D. (2022). The impact of teachers' professional development in science pedagogy on students' achievement: Evidence from TIMSS 2019. *J. Balt. Sci. Educ.* **21** 258–274.

R CORE TEAM (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

STAN DEVELOPMENT TEAM (2023). RStan: The R interface to Stan. R package version 2.32.3.

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. CRC Press, Boca Raton.

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

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

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 329–350
https://doi.org/10.1214/24-AOAS1961

STATISTICAL INFERENCE FOR REGRESSION WITH IMPUTED BINARY COVARIATES WITH APPLICATION TO EMOTION RECOGNITION

BY ZIQIAN LIN[1,a], DANYANG HUANG[2,d], ZIYU XIONG[1,b] AND HANSHENG WANG[1,c]

[1]*Guanghua School of Management, Peking University,* [a]*linziqian@stu.pku.edu.cn,* [b]*2301111050@stu.pku.edu.cn,* [c]*hansheng@gsm.pku.edu.cn*

[2]*Center for Applied Statistics and School of Statistics, Renmin University of China,* [d]*dyhuang89@126.com*

In the flourishing live streaming industry, accurate recognition of streamers' emotions has become a critical research focus with profound implications for audience engagement and content optimization. However, precise emotion coding typically requires manual annotation by trained experts, making it extremely expensive and time-consuming to obtain complete observational data for large-scale studies. Motivated by this challenge in streamer emotion recognition, we develop here a novel imputation method together with a principled statistical inference procedure for analyzing partially observed binary data. Specifically, we assume for each observation an auxiliary feature vector, which is sufficiently cheap to be fully collected for the whole sample. We next assume a small pilot sample with both the target binary covariates (i.e., the emotion status) and the auxiliary features fully observed, of which the size could be considerably smaller than that of the whole sample. Thereafter, a regression model can be constructed for the target binary covariates and the auxiliary features. This enables us to impute the missing binary features using the fully observed auxiliary features for the entire sample. We establish the associated asymptotic theory for principled statistical inference and present extensive simulation experiments, demonstrating the effectiveness and theoretical soundness of our proposed method. Furthermore, we validate our approach using a comprehensive dataset on emotion recognition in live streaming, demonstrating that our imputation method yields smaller standard errors and is more statistically efficient than using pilot data only. Our findings have significant implications for enhancing user experience and optimizing engagement on streaming platforms.

REFERENCES

BARTLETT, J. W., SEAMAN, S. R., WHITE, I. R. and CARPENTER, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat. Methods Med. Res.* **24** 462–487. MR3372102 https://doi.org/10.1177/0962280214521348

CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75.

CHEN, Y.-S., LIU, Z.-Q., CHEN, R., WANG, L., HUANG, L., ZHU, X., ZHOU, T.-Y., LU, W. and MA, P. (2017). Population pharmacokinetic analysis of tacrolimus in Chinese myasthenia gravis patients. *Acta Pharmacol. Sin.* **38** 1195–1204.

DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G. et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. Preprint. Available at arXiv:2010.11929.

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

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099

FANG, F., HONG, Q. and SHAO, J. (2009). A pseudo empirical likelihood approach for stratified samples with nonresponse. *Ann. Statist.* **37** 371–393. MR2488356 https://doi.org/10.1214/07-AOS578

HASSANI, A., WALTON, S., SHAH, N., ABUDUWEILI, A., LI, J. and SHI, H. (2021). Escaping the big data paradigm with compact transformers. Preprint. Available at arXiv:2104.05704.

Key words and phrases. Auxiliary feature, emotion recognition, live streaming, pilot sample, regression imputation.

HENDRYCKS, D. and GIMPEL, K. (2016). Gaussian error linear units (gelus). Preprint. Available at arXiv:1606.08415.

KIEFER, J. (1959). Optimum experimental designs. *J. Roy. Statist. Soc. Ser. B* **21** 272–319. MR0113263

KRIZHEVSKY, A., HINTON, G. et al. (2009). Learning multiple layers of features from tiny images.

LECUN, Y., BOSER, B., DENKER, J., HENDERSON, D., HOWARD, R., HUBBARD, W. and JACKEL, L. (1989). Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* **2**.

LEHMANN, E. L. and CASELLA, G. (2006). *Theory of Point Estimation. Springer Texts in Statistics*. Springer, New York. MR1639875

LI, H., SIMCHI-LEVI, D., WU, M. X. and ZHU, W. (2023). Estimating and exploiting the impact of photo layout: A structural approach. *Manage. Sci.* **69** 5209–5233.

LIN, Z., HUANG, D., XIONG, Z. and WANG, H. (2025). Supplement to "Statistical inference for regression with imputed binary covariates with application to emotion recognition." https://doi.org/10.1214/24-AOAS1961SUPP

LITTLE, R. J. A. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LIU, X., LEE, D. and SRINIVASAN, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *J. Mark. Res.* **56** 918–943.

NILSBACK, M.-E. and ZISSERMAN, A. (2008). Automated flower classification over a large number of classes. In 2008 *Sixth Indian Conference on Computer Vision, Graphics & Image Processing* 722–729. IEEE.

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

PAN, R., REN, T., GUO, B., LI, F., LI, G. and WANG, H. (2022). A note on distributed quantile regression by pilot sampling and one-step updating. *J. Bus. Econom. Statist.* **40** 1691–1700. MR4492062 https://doi.org/10.1080/07350015.2021.1961789

QIAO, X. and LIU, Y. (2009). Adaptive weighted learning for unbalanced multicategory classification. *Biometrics* **65** 159–168. MR2665857 https://doi.org/10.1111/j.1541-0420.2008.01017.x

QIN, J., SHAO, J. and ZHANG, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *J. Amer. Statist. Assoc.* **103** 797–810. MR2524011 https://doi.org/10.1198/016214508000000238

RABINER, L. and SCHAFER, R. (2010). *Theory and Applications of Digital Speech Processing*. Prentice Hall, New York.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0346957

REBUFFI, S.-A., BILEN, H. and VEDALDI, A. (2017). Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems* **30**.

RUSSELL, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* **39** 1161.

SHAO, J. (2003). *Mathematical Statistics*. Springer, Berlin.

SHAO, J. and WANG, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *J. Amer. Statist. Assoc.* **97** 544–552. MR1941471 https://doi.org/10.1198/016214502760047078

SIMONYAN, K. and ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint. Available at arXiv:1409.1556.

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

TIMOSHENKO, A. and HAUSER, J. R. (2019). Identifying customer needs from user-generated content. *Mark. Sci.* **38** 1–20.

WANG, H. (2020). Logistic regression for massive data with rare events. In *International Conference on Machine Learning* 9829–9836. PMLR.

WANG, H., ZHANG, A. and WANG, C. (2021). Nonuniform negative sampling and log odds correction with rare events data. *Adv. Neural Inf. Process. Syst.* **34** 19847–19859.

WANG, H. J. and FENG, X. (2012). Multiple imputation for M-regression with censored covariates. *J. Amer. Statist. Assoc.* **107** 194–204. MR2949351 https://doi.org/10.1080/01621459.2011.643198

WANG, W., HUANG, M., ZHENG, S., LIN, L. and WANG, L. (2022). The impact of broadcasters on consumer's intention to follow livestream brand community. *Front. Psychol.* **12** 810883.

WEI, Y., MA, Y. and CARROLL, R. J. (2012). Multiple imputation in quantile regression. *Biometrika* **99** 423–438. MR2931263 https://doi.org/10.1093/biomet/ass007

WHITE, I. R. and CARLIN, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* **29** 2920–2931. MR2758387 https://doi.org/10.1002/sim.3944

WHITE, I. R. and ROYSTON, P. (2009). Imputing missing covariate values for the Cox model. *Stat. Med.* **28** 1982–1998. MR2750806 https://doi.org/10.1002/sim.3618

ZHANG, M. and LUO, L. (2023). Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Manage. Sci.* **69** 25–50.

ZHANG, S., LEE, D., SINGH, P. V. and SRINIVASAN, K. (2022). What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Manage. Sci.* **68** 5644–5666.

ZHOU, R. and TONG, L. (2022). A study on the influencing factors of consumers' purchase intention during livestreaming e-commerce: The mediating effect of emotion. *Front. Psychol.* **13** 903023.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 351–370
https://doi.org/10.1214/24-AOAS1963

OVERLAP VIOLATIONS IN EXTERNAL VALIDITY: APPLICATION TO UGANDAN CASH TRANSFER PROGRAMS

BY MELODY HUANG[a]

Department of Political Science and Statistics & Data Science, Yale University, [a]*melody.huang@yale.edu*

Estimating externally valid causal effects is a foundational problem in the social and biomedical sciences. Generalizing or transporting causal estimates from an experimental sample to a target population of interest relies on an overlap (or positivity) assumption between the experimental sample and the target population. In practice, having full overlap between an experimental sample and a target population can be implausible. In the following paper, we introduce a framework for considering external validity in the presence of overlap violations. We propose a novel bias decomposition that parameterizes the bias from an overlap violation into two components: (1) the proportion of units omitted and (2) the degree to which omitting the units moderates the treatment effect. The bias decomposition offers an intuitive and straightforward approach to conducting sensitivity analysis to assess robustness to overlap violations. Furthermore, we introduce a suite of sensitivity tools in the form of summary measures and benchmarking, which help researchers consider the plausibility of the overlap violations. We illustrate the proposed framework on an experiment evaluating the impact of a cash transfer program in Northern Uganda.

REFERENCES

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BEN-MICHAEL, E., HIRSCHBERG, D., FELLER, A. and ZUBIZARRETA, J. (2020). The balancing act for causal inference.

BLATTMAN, C., FIALA, N. and MARTINEZ, S. (2013). Generating skilled self-employment in developing countries: Experimental evidence from Uganda. *Q. J. Econ.* **129** 697–752.

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

CHENG, C., LI, F., THOMAS, L. E. and LI, F. (2022). Addressing extreme propensity scores in estimating counterfactual s urvival functions via the overlap weights. *Amer. J. Epidemiol.* **191** 1140–1151.

CINELLI, C. and HAZLETT, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 39–67. MR4060976

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The actg 320 trial. *Amer. J. Epidemiol.* **172** 107–115.

COLNET, B., JOSSE, J., SCORNET, E. and VAROQUAUX, G. (2021). Generalizing a causal effect: Sensitivity analysis and missing covariates. Preprint. Available at arXiv:2105.06435.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

DAHABREH, I. J., ROBERTSON, S. E., STEINGRIMSSON, J. A., STUART, E. A. and HERNÁN, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Stat. Med.* **39** 1999–2014. MR4105273 https://doi.org/10.1002/sim.8426

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

Key words and phrases. Generalizability, transportability, sensitivity analysis, causal inference.

DAHABREH, I. J., ROBINS, J. M., HANEUSE, S. J.-P. A., SAEED, I., ROBERTSON, S. E., STUART, E. A. and HERNÁN, M. A. (2023). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Stat. Med.* **42** 2029–2043. MR4594699 https://doi.org/10.1002/sim.9550

DING, P., FELLER, A. and MIRATRIX, L. (2019). Decomposing treatment effect variation. *J. Amer. Statist. Assoc.* **114** 304–317. MR3941256 https://doi.org/10.1080/01621459.2017.1407322

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.

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

EGAMI, N. and LEE, D. D. I. (2023). Designing multi-context studies for external validity: Site selection via synthetic purposive sampling.

GERBER, A. and GREEN, D. (2012). *Field Experiments: Design, Analysis, and Interpretation*. Norton, New York.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HARTMAN, E. and HUANG, M. (2023). Sensitivity analysis for survey weights. *Polit. Anal.* **32** 1–16.

HONG, G., YANG, F. and QIN, X. (2021). Did you conduct a sensitivity analysis? A new weighting-based approach for evaluations of the average treatment effect for the treated. *J. Roy. Statist. Soc. Ser. A* **184** 227–254. MR4204918 https://doi.org/10.1111/rssa.12621

HUANG, M. (2024). Sensitivity analysis for the generalization of experimental results. *J. R. Stat. Soc., A*. To appear.

HUANG, M. (2025). Supplement to "Overlap violations in external validity: Application to Ugandan cash transfer programs." https://doi.org/10.1214/24-AOAS1963SUPPA, https://doi.org/10.1214/24-AOAS1963SUPPB

HUANG, M. and PIMENTEL, S. (2024). Variance-based sensitivity models for weighted estimators result in more informative bounds. *Biometrika*. To appear.

HUANG, M., SORIANO, D. and PIMENTEL, S. D. (2023). Design sensitivity and its implications for weighted observational studies. Preprint. Available at arXiv:2202.03408.

IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. MR2427345 https://doi.org/10.1111/j.1467-985X.2007.00527.x

JOSEY, K. P., BERKOWITZ, S. A., GHOSH, D. and RAGHAVAN, S. (2021). Transporting experimental results with entropy balancing. *Stat. Med.* **40** 4310–4326. MR4300088 https://doi.org/10.1002/sim.9031

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.

LEI, L., D'AMOUR, A., DING, P., FELLER, A. and SEKHON, J. (2021). Distribution-free assessment of population overlap in observational studies. Technical report, Working paper, Stanford Univ.

LI, F., BUCHANAN, A. L. and COLE, S. R. (2022). Generalizing trial evidence to target populations in non-nested designs: Applications to AIDS clinical trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 669–697. MR4441621 https://doi.org/10.1111/rssc.12550

LU, B., BEN-MICHAEL, E., FELLER, A. and MIRATRIX, L. (2021). Is it who you are or where you are? Accounting for compositional differences in cross-site treatment variation. Preprint. Available at arXiv:2103.14765.

MANSKI, C. F. (1989). Anatomy of the selection problem. *J. Hum. Resour.* **24** 343–360.

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

NIE, X., IMBENS, G. and WAGER, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. Preprint. Available at arXiv:2112.04723.

OLSEN, R. B., ORR, L. L., BELL, S. H. and STUART, E. A. (2013). External validity in policy evaluations that choose sites purposively. *J. Policy Anal. Manage.* **32** 107–121.

PARIKH, H., ROSS, R., STUART, E. and RUDOLPH, K. (2024). Who are we missing? A principled approach to characterizing the underrepresented population. Preprint. Available at arXiv:2401.14512.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537 https://doi.org/10.1177/0962280210386207

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (2003). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

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

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

TIPTON, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* **39** 478–501.

TIPTON, E. (2021). Beyond generalization of the ATE: Designing randomized trials to understand treatment effect heterogeneity. *J. Roy. Statist. Soc. Ser. A* **184** 504–521. MR4255908 https://doi.org/10.1111/rssa.12629

TIPTON, E. and MAMAKOS, M. (2023). Designing randomized experiments to predict unit-specific treatment effects. Preprint. Available at arXiv:2310.18500.

WESTREICH, D. and COLE, S. R. (2010). Invited commentary: Positivity in practice. *Amer. J. Epidemiol.* **171** 674–677.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 371–396
https://doi.org/10.1214/24-AOAS1964

A THREE-STATE COUPLED MARKOV SWITCHING MODEL FOR COVID-19 OUTBREAKS ACROSS QUEBEC BASED ON HOSPITAL ADMISSIONS

BY DIRK DOUWES-SCHULTZ[a], ALEXANDRA M. SCHMIDT[b], YANNAN SHEN[c] AND DAVID L. BUCKERIDGE[d]

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, [a]*dirk.douwes-schultz@mail.mcgill.ca,* [b]*alexandra.schmidt@mcgill.ca,* [c]*yannan.shen@mail.mcgill.ca,* [d]*david.buckeridge@mcgill.ca*

Recurrent COVID-19 outbreaks have placed immense strain on the hospital system in Quebec. We develop a Bayesian three-state coupled Markov switching model to analyze COVID-19 outbreaks across Quebec based on admissions in the 30 largest hospitals. Within each catchment area, we assume the existence of three states for the disease: absence, a new state meant to account for many zeroes in some of the smaller areas; endemic and outbreak. Then we assume the disease switches between the three states in each area through a series of coupled nonhomogeneous hidden Markov chains. Unlike previous approaches, the transition probabilities may depend on covariates and the occurrence of outbreaks in neighboring areas to account for geographical outbreak spread. Additionally, to prevent rapid switching between endemic and outbreak periods we introduce clone states into the model which enforce minimum endemic and outbreak durations. We make some interesting findings, such as that mobility in retail and recreation venues had a positive association with the development and persistence of new COVID-19 outbreaks in Quebec. Based on model comparison, our contributions show promise in improving state estimation retrospectively and in real-time, especially when there are smaller areas and highly spatially synchronized outbreaks. Furthermore, our approach offers new and interesting epidemiological interpretations, such as being able to estimate the effect of covariates on disease extinction.

REFERENCES

AMORÓS, R., CONESA, D., LÓPEZ-QUÍLEZ, A. and MARTINEZ-BENEITO, M.-A. (2020). A spatiotemporal hierarchical Markov switching model for the early detection of influenza outbreaks. *Stoch. Environ. Res. Risk Assess.* **34** 275–292. https://doi.org/10.1007/s00477-020-01773-5

AMORÓS SALVADOR, R. (2017). Bayesian temporal and spatio-temporal Markov switching models for the detection of influenza outbreaks Ph.D. thesis Universitat de València. https://doi.org/10550/59265

AUGER-MÉTHÉ, M., NEWMAN, K., COLE, D., EMPACHER, F., GRYBA, R., KING, A. A., LEOS-BARAJAS, V., MILLS FLEMMING, J., NIELSEN, A. et al. (2021). A guide to state–space modeling of ecological time series. *Ecol. Monogr.* **91** e01470. https://doi.org/10.1002/ecm.1470

BARTLETT, M. S. (1957). Measles periodicity and community size. *J. R. Stat. Soc., A* **120** 48–60. https://doi.org/10.2307/2342553

BAUER, C. and WAKEFIELD, J. (2018). Stratified space-time infectious disease modelling, with an application to hand, foot and mouth disease in China. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 1379–1398. MR3873712 https://doi.org/10.1111/rssc.12284

BAUER, C., WAKEFIELD, J., RUE, H., SELF, S., FENG, Z. and WANG, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Stat. Med.* **35** 1848–1865. MR3513489 https://doi.org/10.1002/sim.6785

BI, S., BROGGI, M. and BEER, M. (2019). The role of the Bhattacharyya distance in stochastic model updating. *Mech. Syst. Signal Process.* **117** 437–452. https://doi.org/10.1016/j.ymssp.2018.08.017

BJØRNSTAD, O. N. (2023). *Epidemics: Models and Data Using R. Use R!* Springer, Cham. https://doi.org/10.1007/978-3-319-97487-3

Key words and phrases. Bayesian inference, COVID-19, hidden Markov model, outbreak detection, outbreak forecasting, zero-inflation.

BRACHER, J. and HELD, L. (2022). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *Int. J. Forecast.* **38** 1221–1233. https://doi.org/10.1016/j.ijforecast.2020.07.002

BUCKERIDGE, D. L. (2007). Outbreak detection through automated surveillance: A review of the determinants of detection. *J. Biomed. Inform.* **40** 370–379. https://doi.org/10.1016/j.jbi.2006.09.003

BULL, S. B., MAK, C. and GREENWOOD, C. M. T. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Comput. Statist. Data Anal.* **39** 57–74. MR1895558 https://doi.org/10.1016/S0167-9473(01)00048-2

CABRERA, H. and OLSON, I. (2021). Facing onset of 4th wave of COVID-19 infections, Quebec to implement vaccine passport system. *CBC News*.

CHENCHULA, S., KARUNAKARAN, P., SHARMA, S. and CHAVAN, M. (2022). Current evidence on efficacy of COVID-19 booster dose vaccination against the Omicron variant: A systematic review. *J. Med. Virol.* **94** 2969–2976. https://doi.org/10.1002/jmv.27697

CHIB, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometrics* **75** 79–97. https://doi.org/10.1016/0304-4076(95)01770-4

CRAWFORD, F. W., JONES, S. A., CARTTER, M., DEAN, S. G., WARREN, J. L., LI, Z. R., BARBIERI, J., CAMPBELL, J., KENNEY, P. et al. (2022). Impact of close interpersonal contact on COVID-19 incidence: Evidence from 1 year of mobile device data. *Sci. Adv.* **8** eabi5499. https://doi.org/10.1126/sciadv.abi5499

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

DESCLOUX, E., MANGEAS, M., MENKES, C. E., LENGAIGNE, M., LEROY, A., TEHEI, T., GUILLAUMOT, L., TEURLAI, M., GOURINAT, A.-C. et al. (2012). Climate–based models for understanding and forecasting dengue epidemics. *PLoS Negl. Trop. Dis.* **6** e1470. https://doi.org/10.1371/journal.pntd.0001470

DIEBOLD, F., LEE, J. and WEINBACH, G. (1994). Regime switching with time–varying transition probabilities. In *Nonstationary Time Series Analysis and Cointegration* (C. Hargreaves, ed.) 283–302. Oxford Univ. Press, Oxford. https://doi.org/10.1515/9780691219585-009

DOUWES-SCHULTZ, D. and SCHMIDT, A. M. (2022). Zero-state coupled Markov switching count models for spatio-temporal infectious disease spread. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 589–612. MR4441618 https://doi.org/10.1111/rssc.12547

DOUWES-SCHULTZ, D., SCHMIDT, A. M., BUCKERIDGE, D. and SHEN, Y. (2025). Supplement to "A three-state coupled Markov switching model for COVID-19 outbreaks across Quebec based on hospital admissions." https://doi.org/10.1214/24-AOAS1964SUPPA, https://doi.org/10.1214/24-AOAS1964SUPPB

FLAXMAN, S., MISHRA, S., GANDY, A., UNWIN, H. J. T., MELLAN, T. A., COUPLAND, H., WHITTAKER, C., ZHU, H., BERAH, T. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** 257–261. https://doi.org/10.1038/s41586-020-2405-7

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*. Springer, New York. MR2265601 https://doi.org/10.1007/978-0-387-35768-3

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

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GILMOUR, S. J. (2010). Identification of hospital catchment areas using clustering: An example from the NHS. *Health Serv. Res.* **45** 497–513. https://doi.org/10.1111/j.1475-6773.2009.01069.x

GOOGLE LLC (2022). COVID-19 community mobility report. https://www.google.com/covid19/mobility?hl=en.

GOUVERNEMENT DU QUÉBEC (2022). Measures adopted by orders in council and ministerial orders in the context of the COVID-19 pandemic. https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/measures-orders-in-council-ministerial-orders.

GRENFELL, B. T., BJØRNSTAD, O. N. and KAPPEY, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414** 716–723. https://doi.org/10.1038/414716a

HAMILTON, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57** 357–384. MR0996941 https://doi.org/10.2307/1912559

HEATON, M. J., BANKS, D. L., ZOU, J., KARR, A. F., DATTA, G., LYNCH, J. and VERA, F. (2012). A spatio-temporal absorbing state model for disease and syndromic surveillance. *Stat. Med.* **31** 2123–2136. MR2956066 https://doi.org/10.1002/sim.5350

INSPQ (2022). COVID-19 data in Quebec [online]. https://www.inspq.qc.ca/covid-19/donnees.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987 https://doi.org/10.1214/088342305000000016

JOHNSON, A. G., AMIN, A. B., ALI, A. R., HOOTS, B., CADWELL, B. L., ARORA, S., AVOUNDJIAN, T., AWOFESO, A. O., BARNES, J. et al. (2022). COVID-19 incidence and death rates among unvaccinated and fully vaccinated adults with and without booster doses during periods of Delta and Omicron variant emergence - 25 U.S. jurisdictions, April 4-December 25, 2021. *Morb. Mort. Wkly. Rep.* **71** 132–138. https://doi.org/10.15585/mmwr.mm7104e2

KAUFMANN, S. (2018). Hidden Markov models in time series, with applications in economics. In *Handbook of Mixture Analysis* (S. Frühwirth-Schnatter, G. Celeux and C. P. Robert, eds.) 309–341. CRC Press/CRC, New York.

KEELING, M. J. and ROHANI, P. (2007). *Modeling Infectious Diseases in Humans and Animals*, Illustrated edition. Princeton Univ. Press, Princeton, NJ. https://doi.org/10.2307/j.ctvcm4gk0

KNORR-HELD, L. and RICHARDSON, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *J. R. Stat. Soc., Ser. C* **52** 169–183. MR1973739 https://doi.org/10.1111/1467-9876.00396

LAFRAMBOISE, K. (2020). Quebec coronavirus cases soar past 4,000 as medical equipment shortage looms. *Glob. News*.

LANGROCK, R. and ZUCCHINI, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Comput. Statist. Data Anal.* **55** 715–724. MR2736590 https://doi.org/10.1016/j.csda.2010.06.015

LEGAULT, F. and BLAIS, M. (2020). Conférence de presse de M. François Legault, premier ministre et Mme Marguerite Blais, ministre responsable des aînés et des proches aidants, Quebec City. http://www.assnat.qc.ca/fr/actualites-salle-presse/conferences-points-presse/ConferencePointPresse-60025.html.

LIBOSCHIK, T., FOKIANOS, K. and FRIED, R. (2017). Tscount: An R package for analysis of count time series following generalized linear models. *J. Stat. Softw.* **82** 1–51. https://doi.org/10.18637/jss.v082.i05

LU, H.-M., ZENG, D. and CHEN, H. (2010). Prospective infectious disease outbreak detection using Markov switching models. *IEEE Trans. Knowl. Data Eng.* **22** 565–577. https://doi.org/10.1109/TKDE.2009.115

LYTRAS, T., GKOLFINOPOULOU, K., BONOVAS, S. and NUNES, B. (2019). FluHMM: A simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Stat. Methods Med. Res.* **28** 1826–1840. MR3961968 https://doi.org/10.1177/0962280218776685

MARTÍNEZ-BENEITO, M. A., CONESA, D., LÓPEZ-QUÍLEZ, A. and LÓPEZ-MASIDE, A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics. *Stat. Med.* **27** 4455–4468. MR2528524 https://doi.org/10.1002/sim.3320

MASLO, C., FRIEDLAND, R., TOUBKIN, M., LAUBSCHER, A., AKALOO, T. and KAMA, B. (2022). Characteristics and outcomes of hospitalized patients in South Africa during the COVID-19 Omicron wave compared with previous waves. *JAMA* **327** 583–584. https://doi.org/10.1001/jama.2021.24868

MIZUMOTO, K. and CHOWELL, G. (2020). Transmission potential of the novel coronavirus (COVID-19) onboard the diamond Princess Cruises Ship, 2020. *Infect. Dis. Model.* **5** 264–270. https://doi.org/10.1016/j.idm.2020.02.003

NCCID (2022). Updates on COVID-19 variants of concern (VOC). https://nccid.ca/covid-19-variants/.

NUNES, B., NATÁRIO, I. and CARVALHO, M. L. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Stat. Med.* **32** 2643–2660. MR3067413 https://doi.org/10.1002/sim.5670

OLIVIER, A. (2021). Quebec imposes lockdown for 3 cities as COVID-19 cases rise ahead of Easter long weekend. *Glob. News*.

PAUL, M. and HELD, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Stat. Med.* **30** 1118–1136. MR2767846 https://doi.org/10.1002/sim.4177

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

POHLE, J., LANGROCK, R., VAN DER SCHAAR, M., KING, R. and JENSEN, F. H. (2021). A primer on coupled state-switching models for multiple interacting time series. *Stat. Model.* **21** 264–285. MR4249702 https://doi.org/10.1177/1471082X20956423

RAHIMI, I., CHEN, F. and GANDOMI, A. H. (2021). A review on COVID-19 forecasting models. *Neural Comput. Appl.* **35** 23671–23681. https://doi.org/10.1007/s00521-020-05626-8

RAHMANIAN, V., BOKAIE, S., HAGHDOOST, A. and BAROUNI, M. (2021). Predicting cutaneous leishmaniasis using SARIMA and Markov switching models in Isfahan, Iran: A time-series study. *Asian Pac. J. Trop. Med.* **14** 83–93. https://doi.org/10.4103/1995-7645.306739

RATH, T., CARRERAS, M. and SEBASTIANI, P. (2003). Automated detection of influenza epidemics with hidden Markov models. In *Advances in Intelligent Data Analysis V. IDA* 2003 (M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse and C. Borgelt, eds.). *Lecture Notes in Computer Science* **2810** 521–532. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-45231-7_48

SCHRÖDLE, B., HELD, L. and RUE, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics* **68** 736–744. MR3055178 https://doi.org/10.1111/j.1541-0420.2011.01717.x

SHABY, B. A. and WELLS, M. T. (2010). Exploring an adaptive Metropolis algorithm.

SHINGLER, B. and HENDRY, L. (2022). Overwhelmed by COVID, Quebec hospitals face tough choices in scaling back surgeries. *CBC News*.

SOPER, H. E. (1929). The interpretation of periodicity in disease prevalence. *J. R. Stat. Soc.* **92** 34–73. https://doi.org/10.2307/2341437

SOUZA, R. C. (1982). Forecasting the progress of epidemics by means of a Bayesian-entropy framework. *Environ. Plann. A, Econ. Space* **14** 49–60. https://doi.org/10.1068/a140049

SSENTONGO, P., FRONTERRE, C., GERONIMO, A., GREYBUSH, S. J., MBABAZI, P. K., MUVAWALA, J., NAHALAMBA, S. B., OMADI, P. O., OPAR, B. T. et al. (2021). Pan-African evolution of within- and between-country COVID-19 dynamics. *Proc. Natl. Acad. Sci. USA* **118** e2026664118. https://doi.org/10.1073/pnas.2026664118

STEVENSON, V. (2021). Quebec shuts down schools, bars, gyms and more as COVID-19 case counts soar. *CBC News*.

STOJANOVIĆ, O., LEUGERING, J., PIPA, G., GHOZZI, S. and ULLRICH, A. (2019). A Bayesian Monte Carlo approach for predicting the spread of infectious diseases. *PLoS ONE* **14** e0225838. https://doi.org/10.1371/journal.pone.0225838

STONER, O. and ECONOMOU, T. (2020). An advanced hidden Markov model for hourly rainfall time series. *Comput. Statist. Data Anal.* **152** 107045. MR4126383 https://doi.org/10.1016/j.csda.2020.107045

TOULOUPOU, P., FINKENSTÄDT, B. and SPENCER, S. E. F. (2020). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *J. Comput. Graph. Statist.* **29** 238–249. MR4116038 https://doi.org/10.1080/10618600.2019.1654880

UNKEL, S., FARRINGTON, C. P., GARTHWAITE, P. H., ROBERTSON, C. and ANDREWS, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *J. Roy. Statist. Soc. Ser. A* **175** 49–82. MR2873791 https://doi.org/10.1111/j.1467-985X.2011.00714.x

VERELST, F., WILLEM, L. and BEUTELS, P. (2016). Behavioural change models for infectious disease transmission: A systematic review (2010-2015). *J. R. Soc. Interface* **13**. https://doi.org/10.1098/rsif.2016.0820

WAKEFIELD, J. (2013). *Bayesian and Frequentist Regression Methods. Springer Series in Statistics*. Springer, New York. MR3025225 https://doi.org/10.1007/978-1-4419-0925-1

WAKEFIELD, J., DONG, T. Q. and MININ, V. N. (2019). Spatio-temporal analysis of surveillance data. In *Handbook of Infectious Disease Data Analysis* (L. Held, N. Hens, P. O'Neil and J. Wallinga, eds.) 455–475. CRC Press/CRC, Boca Raton, FL.

WANG, E. T., CHIANG, S., HANEEF, Z., RAO, V. R., MOSS, R. and VANNUCCI, M. (2023). Bayesian non-homogeneous hidden Markov model with variable selection for investigating drivers of seizure risk cycling. *Ann. Appl. Stat.* **17** 333–356. MR4539034 https://doi.org/10.1214/22-aoas1630

WARD, T. and JOHNSEN, A. (2021). Understanding an evolving pandemic: An analysis of the clinical time delay distributions of COVID-19 in the United Kingdom. *PLoS ONE* **16** e0257978. https://doi.org/10.1371/journal.pone.0257978

WATKINS, R. E., EAGLESON, S., VEENENDAAL, B., WRIGHT, G. and PLANT, A. J. (2009). Disease surveillance using a hidden Markov model. *BMC Med. Inform. Decis. Mak.* **9** 39. https://doi.org/10.1186/1472-6947-9-39

WILTON, K. (2020). COVID-19 updates March 15: Quebec now has 39 cases, orders further closures. *Montr. Gaz.*

ZACHER, B. and CZOGIEL, I. (2022). Supervised learning using routine surveillance data improves outbreak detection of Salmonella and Campylobacter infections in Germany. *PLoS ONE* **17** e0267510. https://doi.org/10.1371/journal.pone.0267510

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 397–418
https://doi.org/10.1214/24-AOAS1965

DISENTANGLING THE STRUCTURE OF ECOLOGICAL BIPARTITE NETWORKS FROM OBSERVATION PROCESSES

BY EMRE ANAKOK[1,a], PIERRE BARBILLON[1,b], COLIN FONTAINE[2,c] AND ELISA THEBAULT[3,d]

[1]*Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay,* [a]*emre.anakok@agroparistech.fr,* [b]*pierre.barbillon@agroparistech.fr*

[2]*Centre d'Écologie et des Sciences de la Conservation, MNHN, CNRS, SU,* [c]*colin.fontaine@mnhn.fr*

[3]*Sorbonne Université, CNRS, IRD, INRAE, Université Paris Est Créteil, Université Paris Cité, Institute of Ecology and Environmental Sciences (iEES-Paris),* [d]*elisa.thebault@upmc.fr*

The structure of a bipartite interaction network can be described by providing a clustering for each of the two types of nodes. Such clusterings are outputted by fitting a latent block model (LBM) on an observed network that comes from a sampling of species interactions. However, the sampling is limited and possibly uneven. This may jeopardize the fit of the LBM and then the description of the structure of the network by detecting structures resulting from the sampling and not from actual underlying ecological phenomena. If the observed interaction network consists of a weighted bipartite network where the number of observed interactions between two species is available, the sampling efforts for all species can be estimated and used to correct the LBM fit. We propose to combine an observation model that accounts for sampling and an LBM for describing the structure of underlying possible ecological interactions. We develop an original inference procedure for this model, the efficiency of which is demonstrated in simulation studies. Its relevance and its practical interest are attested on a large dataset of plant-pollinator networks, as we observe structural change on most of the networks depending on whether observation processes are accounted for or not.

REFERENCES

AICHER, C., JACOBS, A. Z. and CLAUSET, A. (2015). Learning latent block structure in weighted networks. *J. Complex Netw.* **3** 221–248. MR3365464 https://doi.org/10.1093/comnet/cnu026

ALMEIDA-NETO, M., GUIMARÃES, P., GUIMARÃES JR, P. R., LOYOLA, R. D. and ULRICH, W. (2008). A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos* **117** 1227–1239.

ANAKOK, E., BARBILLON, P., FONTAINE, C. and THEBAULT, E. (2025). Supplement to "Disentangling the structure of ecological bipartite networks from observation processes." https://doi.org/10.1214/24-AOAS1965SUPP

BALDRIDGE, E., HARRIS, D. J., XIAO, X. and WHITE, E. P. (2016). An extensive comparison of species-abundance distribution models. *PeerJ* **4** e2823. https://doi.org/10.7717/peerj.2823

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

BLÜTHGEN, N., FRÜND, J., VÁZQUEZ, D. P. and MENZEL, F. (2008). What do interaction network metrics tell us about specialization and biological traits? *Ecology* **89** 3387–3399. https://doi.org/10.1890/07-2121.1

BLÜTHGEN, N., STORK, N. and FIEDLER, K. (2004). Bottom-up control and co-occurrence in complex communities: Honeydew and nectar determine a rainforest ant mosaic. *Oikos* **106** 344–358.

BRAULT, V., KERIBIN, C., CELEUX, G. and GOVAERT, G. (2014). Estimation and selection for the latent block model on categorical data. *Stat. Comput.* **25** 1–16.

CELEUX, G. and DIEBOLT, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stoch. Stoch. Rep.* **41** 119–134. MR1275369 https://doi.org/10.1080/17442509208833797

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

Key words and phrases. Latent block model, modularity, nestedness, sampling effect, stochastic expectation maximization.

CHACOFF, N. P., VÁZQUEZ, D. P., LOMÁSCOLO, S. B., STEVANI, E. L., DORADO, J. and PADRÓN, B. (2012). Evaluating sampling completeness in a desert plant-pollinator network: Sampling a plant-pollinator network. *J. Anim. Ecol.* **81** 190–200.

CHAMBERLAIN, S. A., CARTAR, R. V., WORLEY, A. C., SEMMLER, S. J., GIELENS, G., ELWELL, S., EVANS, M. E., VAMOSI, J. C. and ELLE, E. (2014). Traits and phylogenetic history contribute to network structure across Canadian plant–pollinator communities. *Oecologia* **176** 545–556.

CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **11** 265–270. MR0793175

CHAO, A. and JOST, L. (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology* **93** 2533–2547. https://doi.org/10.1890/11-1952.1

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

DE MANINCOR, N., HAUTEKÈETE, N., MAZOYER, C., MOREAU, P., PIQUOT, Y., SCHATZ, B., SCHMITT, E., ZÉLAZNY, M. and MASSOL, F. (2020). How biased is our perception of plant-pollinator networks? A comparison of visit- and pollen-based representations of the same networks. *Acta Oecol.* **105** 103551.

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103 https://doi.org/10.1214/aos/1018031103

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

DORÉ, M., FONTAINE, C. and THÉBAULT, E. (2021). Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Glob. Change Biol.* **27** 1266–1280. https://doi.org/10.1111/gcb.15474

DOSER, J. W., FINLEY, A. O., KÉRY, M. and ZIPKIN, E. F. (2022a). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods Ecol. Evol.* **13** 1670–1678.

DOSER, J. W., LEUENBERGER, W., SILLETT, T. S., HALLWORTH, M. T. and ZIPKIN, E. F. (2022b). Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods Ecol. Evol.* **13** 919–932.

DURAND-BESSART, C., CORDEIRO, N. J., CHAPMAN, C. A., ABERNETHY, K., FORGET, P.-M., FONTAINE, C. and BRETAGNOLLE, F. (2023). Trait matching and sampling effort shape the structure of the frugivory network in Afrotropical forests. *New Phytol.* **237** 1446–1462. https://doi.org/10.1111/nph.18619

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. MR0120167 https://doi.org/10.5486/pmd.1959.6.3-4.12

FORT, H., VÁZQUEZ, D. P. and LAN, B. L. (2016). Abundance and generalisation in mutualistic networks: Solving the chicken-and-egg dilemma. *Ecol. Lett.* **19** 4–11.

FORTUNA, M. A., STOUFFER, D. B., OLESEN, J. M., JORDANO, P., MOUILLOT, D., KRASNOV, B. R., POULIN, R. and BASCOMPTE, J. (2010). Nestedness versus modularity in ecological networks: Two sides of the same coin? *J. Anim. Ecol.* **79** 811–817. https://doi.org/10.1111/j.1365-2656.2010.01688.x

GOVAERT, G. and NADIF, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches. *Comput. Statist. Data Anal.* **52** 3233–3245. MR2424788 https://doi.org/10.1016/j.csda.2007.09.007

GOVAERT, G. and NADIF, M. (2010). Latent block model for contingency table. *Comm. Statist. Theory Methods* **39** 416–425. MR2745285 https://doi.org/10.1080/03610920903140197

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

HADFIELD, J. D., KRASNOV, B. R., POULIN, R. and NAKAGAWA, S. (2014). A tale of two phylogenies: Comparative analyses of ecological interactions. *Amer. Nat.* **183** 174–187. https://doi.org/10.1086/674445

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

INGS, T. C. M. et al. (2009). Review: Ecological networks – beyond food webs. *J. Anim. Ecol.* **78** 253–269.

JORDANO, P. (2016). Sampling networks of ecological interactions. *Funct. Ecol.* **30** 1883–1893.

KAISER-BUNBURY, C. N., MUFF, S., MEMMOTT, J., MÜLLER, C. B. and CAFLISCH, A. (2010). The robustness of pollination networks to the loss of species and interactions: A quantitative approach incorporating pollinator behaviour. *Ecol. Lett.* **13** 442–452.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

KÉFI, S., MIELE, V., WIETERS, E. A., NAVARRETE, S. A. and BERLOW, E. L. (2016). How structured is the entangled bank? The surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. *PLoS Biol.* **14** 1–21.

KRISHNA, A., GUIMARÃES JR, P. R., JORDANO, P. and BASCOMPTE, J. (2008). A neutral-niche theory of nestedness in mutualistic networks. *Oikos* **117** 1609–1618.

KUHN, E., MATIAS, C. and REBAFKA, T. (2020). Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Stat*. *Comput*. **30** 1725–1739. MR4156345 https://doi.org/10.1007/s11222-020-09968-0

LARA-ROMERO, C., SEGUÍ, J., PÉREZ-DELGADO, A., NOGALES, M. and TRAVESET, A. (2019). Beta diversity and specialization in plant–pollinator networks along an elevational gradient. *J. Biogeogr.* **46** 1598–1610.

LÉGER, J.-B., DAUDIN, J.-J. and VACHER, C. (2015). Clustering methods differ in their ability to detect patterns in ecological networks. *Methods Ecol*. *Evol*. **6** 474–481.

MACGREGOR, C., EVANS, D. and POCOCK, M. (2017). *Estimating Sampling Completeness of Interactions in Quantitative Bipartite Ecological Networks*: *Incorporating Variation in Species' Specialisation*.

MACKENZIE, D., NICHOLS, J., LACHMAN, G., DROEGE, S., ROYLE, J. A. and LANGTIMM, C. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83** 2248–2255.

MAGRACH, A., HOLZSCHUH, A., BARTOMEUS, I., RIEDINGER, V., ROBERTS, S., RUNDLÖF, M., VUJIC, A., WICKENS, J., WICKENS, V. et al. (2017). Plant-pollinator networks in semi-natural grasslands are resistant to the loss of pollinators during blooming of mass-flowering crops. *Ecography* **41** 62–74.

O'CONNOR, L. M. J., POLLOCK, L. J., BRAGA, J., FICETOLA, G. F., MAIORANO, L., MARTINEZ-ALMOYNA, C., MONTEMAGGIORI, A., OHLMANN, M. and THUILLER, W. (2020). Unveiling the food webs of tetrapods across Europe through the prism of the Eltonian niche. *J. Biogeogr.* **47** 181–192.

OLESEN, J. M., BASCOMPTE, J., DUPONT, Y. L., ELBERLING, H., RASMUSSEN, C. and JORDANO, P. (2011). Missing and forbidden links in mutualistic networks. *Proc*. *Royal Soc*. *B*, *Biol*. *Sci*. **278** 725–732.

OLESEN, J. M., ESKILDSEN, L. I. and VENKATASAMY, S. (2002). Invasion of pollination networks on Oceanic islands: Importance of invader complexes and endemic super generalists. *Divers*. *Distrib.* **8** 181–192.

PAPADOGEORGOU, G., BELLO, C., OVASKAINEN, O. and DUNSON, D. B. (2023). Covariate-informed latent interaction models: Addressing geographic & taxonomic bias in predicting bird-plant interactions. *J. Amer. Statist*. *Assoc*. **118** 2250–2261. MR4681580 https://doi.org/10.1080/01621459.2023.2208390

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist*. *Assoc*. **66** 846–850.

RIVERA-HUTINEL, A., BUSTAMANTE, R. O., MARÍN, V. H. and MEDEL, R. (2012). Effects of sampling completeness on the structure of plant-pollinator networks. *Ecology* **93** 1593–1603. https://doi.org/10.1890/11-1803.1

SANDER, E. L., WOOTTON, J. T. and ALLESINA, S. (2015). What can interaction webs tell us about species roles? *PLoS Comput*. *Biol*. **11** 1–22.

STANICZENKO, P. et al. (2013). The ghost of nestedness in ecological networks. *Nat*. *Commun*. **4** 1391.

SUGIHARA, G. (1980). Minimal community structure: An explanation of species abundance patterns. *Amer. Nat*. **116** 770–787. MR0642397 https://doi.org/10.1086/283669

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

TERRY, J. C. D. and LEWIS, O. T. (2020). Finding missing links in interaction networks. *Ecology* **101** e03047. https://doi.org/10.1002/ecy.3047

ZHAO, Y., HAO, N. and ZHU, J. (2024). Variational estimators of the degree-corrected latent block model for bipartite networks. *J*. *Mach*. *Learn*. *Res*. **25** 150. MR4758121

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 419–439
https://doi.org/10.1214/24-AOAS1966
This research was funded, in whole or in part, by [ESPRC & UKRI, EP/R513283/1, MR/X034992/1 and ESRC, ES/V015370/1]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions

COMPARATIVE JUDGEMENT MODELLING TO MAP FORCED MARRIAGE AT LOCAL LEVELS

BY ROWLAND SEYMOUR[1,a], ALBERT NYARKO-AGYEI[2,b], HELEN MCCABE[2,c], KATIE SEVERN[3,d], DAVID SIRL[3,e], THEODORE KYPRAIOS[3,f] AND ADAM TAYLOR[4,g]

[1]*School of Mathematics, University of Birmingham,* [a]*r.g.seymour@bham.ac.uk*
[2]*Rights Lab, University of Nottingham,* [b]*Albert.Nyarko-Agyei@nottingham.ac.uk,* [c]*Helen.McCabe@nottingham.ac.uk*
[3]*School of Mathematical Sciences, University of Nottingham,* [d]*Katie.Severn@nottingham.ac.uk,* [e]*David.Sirl@nottingham.ac.uk,* [f]*Theodore.Kypraios@nottingham.ac.uk*
[4]*Digital Research Services, University of Nottingham,* [g]*Adam.Taylor@nottingham.ac.uk*

Forcing someone into marriage against their will is a violation of their human rights. In 2021, the county of Nottinghamshire, U.K., launched a strategy to tackle forced marriage and violence against women and girls. We set out to map the risk of forced marriage across the county to support the strategy and enable the development of local interventions. However, there was no centralised database for forced marriage in the county and carrying out a survey using standard survey methods was unlikely to produce robust results due to the hidden nature of this crime. Comparative judgement provides a survey design that can map the risk of forced marriage through pairwise comparisons. Current comparative judgement models require studies to have a large number of participants, so we developed a more flexible spatial modelling structure and a mechanism to schedule comparisons more effectively. The proposed modelling structure reduced the data collection burden and made a comparative judgement study feasible with a small number of participants. Underpinning this structure is a latent variable representation that improves on the scalability of inferential procedures of previous comparative judgement models. We used these methods to map the risk of forced marriage across Nottinghamshire, thereby supporting the county's strategy for tackling violence against women and girls.

REFERENCES

BISSON, M.-J., GILMORE, C., INGLIS, M. and JONES, I. (2016). Measuring conceptual understanding using comparative judgement. *Int. J. Res. Undergrad. Math. Educ.* **2** 141–164.

BLEI, D. M. and FRAZIER, P. I. (2011). Distance dependent Chinese restaurant processes. *J. Mach. Learn. Res.* **12** 2461–2488. MR2834504

BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925 https://doi.org/10.2307/2334029

CARON, F. and DOUCET, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *J. Comput. Graph. Statist.* **21** 174–196. MR2913362 https://doi.org/10.1080/10618600.2012.638220

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

CHANTLER, K., MIRZA, N. and MACKENZIE, M. (2021). Policy and professional responses to forced marriage in Scotland. *British J. Soc. Work* **52** 833–849.

DEVROYE, L. (1986). *Nonuniform Random Variate Generation*. Springer, New York. MR0836973 https://doi.org/10.1007/978-1-4613-8643-8

ESTRADA, E. and HIGHAM, D. J. (2010). Network properties revealed through matrix functions. *SIAM Rev.* **52** 696–714. MR2736969 https://doi.org/10.1137/090761070

GHOSH, S., UNGUREANU, A., SUDDERTH, E. and BLEI, D. (2011). Spatial distance dependent Chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds.) **24**. Curran Associates, Red Hook.

Key words and phrases. Bradley–Terry Model, Bayesian computation, Bayesian nonparametrics, violence against women and girls.

GLICKMAN, M. E. and JENSEN, S. T. (2005). Adaptive paired comparison design. *J. Statist. Plann. Inference* **127** 279–293. MR2103038 https://doi.org/10.1016/j.jspi.2003.09.022

GRASSHOFF, U. and SCHWABE, R. (2008). Optimal design for the Bradley–Terry paired comparison model. *Stat. Methods Appl.* **17** 275–289. MR2425186 https://doi.org/10.1007/s10260-007-0058-4

GRAY, A., RAHAT, A., CRICK, T. and LINDSAY, S. (2024). A Bayesian active learning approach to comparative judgement within education assessment. *Comput. Educ.: Artif. Intell.* 100245.

GUO, Y., TIAN, P., KALPATHY-CRAMER, J., OSTMO, S., CAMPBELL, J. P., CHIANG, M. F., ERDOGMUS, D., DY, J. G. and IOANNIDIS, S. (2018). Experimental design under the Bradley–Terry model. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* 2198–2204.

HC DEB (2022a). 45683. Available at https://questions-statements.parliament.uk/written-questions/detail/2022-09-02/45683.

HC DEB (2022b). 61026. Available at https://questions-statements.parliament.uk/written-questions/detail/2022-10-11/61026.

HOME AFFAIRS COMMMITTEE (2008). Domestic violence, forced marriage and “honour”- based violence: Further government response to the committee’s sixth report of session 2007–08 Technical Report.

HOUSE OF LORDS MODERN SLAVERY ACT 2015 COMMITTEE (2024). Corrected oral evidence 15 April 2024 Technical Report.

JONES, I., BISSON, M., GILMORE, C. and INGLIS, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educ. Res. J.* **45** 662–680.

KHAN, T. and MIKUSKA, É. (2021). The first three weeks of lockdown in England: The challenges of detecting safeguarding issues amid nursery and primary school closures due to COVID-19. *Soc. Sci. Humanit. Open* **3** 100099. https://doi.org/10.1016/j.ssaho.2020.100099

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis. Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Academic Press, London. MR0560319

MARSHALL, H. (2020). Understanding demand for songbirds within Indonesia’s captive bird trade Ph.D. thesis Manchester Metropolitan Univ. Manchester.

MCCABE, H. R., SEYMOUR, R. G., SCHWARZ, K., EGLEN, L. and BROWN, R. (2022). Written evidence HBA0016 Technical Report Women and Equalities Select Committee.

NOTTINGHAMSHIRE SAFEGUARDING CHILDREN PARTNERHSIP (2024). May 2024 Newsletter. Available at https://nscp.nottinghamshire.gov.uk/media/a3njwr2b/202405nscpnewsletter.pdf Accessed: 2024-08-12.

OFFICE OF THE POLICE AND CRIME COMMISSIONER IN NOTTINGHAMSHIRE (2021). Violence against women and girls strategy, Nottingham and Nottinghamshire 2021-2025. Technical Report.

OHISHI, M., FUKUI, K., OKAMURA, K., ITOH, Y. and YANAGIHARA, H. (2019). Estimation for spatial effects by using the fused lasso. *Hiroshima Stat. Res. Group* **19**.

POLLITT, A. (2012). The method of adaptive comparative judgement. *Asses. Educ.: Princi. Pol. Pract.* **19** 281–300.

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

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RYAN, E. G., DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.* **84** 128–154. MR3491282 https://doi.org/10.1111/insr.12107

SEYMOUR, R., NYARKO-AGYEI, A., MCCABE, H., SEVERN, K., SIRL, D., KYPRAIOS, T. and TAYLOR, A. (2025). Supplement to “Comparative judgement modelling to map forced marriage at local levels.” https://doi.org/10.1214/24-AOAS1966SUPPA, https://doi.org/10.1214/24-AOAS1966SUPPB, https://doi.org/10.1214/24-AOAS1966SUPPC

SEYMOUR, R. G. and MCCABE, H. R. (2022). Written evidence SVA0006 Technical Report Committee of Public Accounts.

SEYMOUR, R. G., SIRL, D., PRESTON, S. and GOULDING, J. (2023). Multi-level spatial comparative judgement models to map deprivation. *Proc. Joint Stat. Meet.* 2023.

SEYMOUR, R. G., SIRL, D., PRESTON, S. P., DRYDEN, I. L., ELLIS, M. J. A., PERRAT, B. and GOULDING, J. (2022). The Bayesian spatial Bradley–Terry model: Urban deprivation modelling in Tanzania. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 288–308. MR4396910 https://doi.org/10.1111/rssc.12532

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

TURNER, H. and FIRTH, D. (2012). Bradley–Terry models in R: The BradleyTerry2 package. *J. Stat. Softw.* **48**.

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99** 673–686. MR2090902 https://doi.org/10.1198/016214504000000980

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 440–464
https://doi.org/10.1214/24-AOAS1967

MORE POWER TO YOU: USING MACHINE LEARNING TO AUGMENT HUMAN CODING FOR MORE EFFICIENT INFERENCE IN TEXT-BASED RANDOMIZED TRIALS

BY REAGAN MOZER[1,a] AND LUKE MIRATRIX[2,b]

[1]*Department of Mathematical Sciences, Bentley University,* [a]*rmozer@bentley.edu*
[2]*Graduate School of Education, Harvard University,* [b]*lmiratrix@g.harvard.edu*

For randomized trials that use text as an outcome, traditional approaches for assessing treatment impact require that each document first be manually coded for constructs of interest by trained human raters. This process, the current standard, is both time-consuming and limiting: even the largest human coding efforts are typically constrained to measure only a small set of dimensions across a subsample of available texts. In this work we present an inferential framework that can be used to increase the power of an impact assessment, given a fixed human-coding budget, by taking advantage of any "untapped" observations—those documents not manually scored due to time or resource constraints—as a supplementary resource. Our approach, a methodological combination of causal inference, survey sampling methods, and machine learning, has four steps: (1) select and code a sample of documents, (2) build a machine learning model to predict the human-coded outcomes from a set of automatically extracted text features, (3) generate machine-predicted scores for all documents and use these scores to estimate treatment impacts, and (4) adjust the final impact estimates using the residual differences between human-coded and machine-predicted outcomes. This final step ensures any biases in the modeling procedure do not propagate to biases in final estimated effects. Through an extensive simulation study and an application to a recent field trial in education, we show that our proposed approach can be used to reduce the scope of a human-coding effort while maintaining nominal power to detect a significant treatment impact.

REFERENCES

AMORIM, E., CANÇADO, M. and VELOSO, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume* 1 (*Long Papers*) 229–237.

ANAYA, A. R. and BOTICARIO, J. G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *Expert Syst*. *Appl*. **38** 1171–1181.

ANGWIN, J., LARSON, J., MATTU, S. and KIRCHNER, L. (2016). Machine bias. *ProPublica* **23** 139–159.

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

AZAHAR, M. Z. A. B. and GHAUTH, K. I. B. (2022). A hybrid automated essay scoring using NLP and random forest regression. In *International Conference on Computer*, *Information Technology and Intelligent Computing* 448–457. Atlantis Press.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

NATIONAL GOVERNORS ASSOCIATION CENTER FOR BEST PRACTICES, C. O. C. S. S. O. (2010). *Common Core State Standards Technical Report*, *National Governors Association Center for Best Practices*. Council of Chief State School Officers, Washington, D.C.

BLOOM, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Eval*. *Rev*. **19** 547–556.

BOYD, R. L., ASHOKKUMAR, A., SERAJ, S. and PENNEBAKER, J. W. (2022). In The Development and Psychometric Properties of LIWC-22 1–47. University of Texas at Austin, Austin, TX.

Key words and phrases. Causal inference, automated scoring, randomized experiments, NLP.

BREIDT, F. J. and OPSOMER, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statist*. *Sci*. **32** 190–205. MR3648955 https://doi.org/10.1214/16-STS589

BREIMAN, L. (2017). Classification and regression trees. Routledge.

BURSTEIN, J., KUKICH, K., WOLFF, S., LU, C. and CHODOROW, M. (1998). Computer analysis of essays. In NCME Symposium on automated Scoring.

CALISKAN, A., BRYSON, J. J. and NARAYANAN, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* **356** 183–186. https://doi.org/10.1126/science.aal4230

CHOULDECHOVA, A. and ROTH, A. (2018). The frontiers of fairness in machine learning. Preprint. Available at arXiv:1810.08810.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum, Hillsdale.

CROSSLEY, S. A., BAFFOUR, P., TIAN, Y., FRANKLIN, A., BENNER, M. and BOSER, U. (2023). A large-scale corpus for assessing written argumentation: PERSUADE 2.0.

CROSSLEY, S. A., KYLE, K. and DASCALU, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behav*. *Res*. *Methods* **51** 14–27. https://doi.org/10.3758/s13428-018-1142-4

CROSSLEY, S. A., KYLE, K. and MCNAMARA, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behav*. *Res*. *Methods* **48** 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

DAGDOUG, M., GOGA, C. and HAZIZA, D. (2023). Model-assisted estimation in high-dimensional settings for survey data. *J*. *Appl*. *Stat*. **50** 761–785. MR4545057 https://doi.org/10.1080/02664763.2022.2047905

DENNY, M. J. and SPIRLING, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Polit*. *Anal*. **26** 168–189.

DING, P., LI, X. and MIRATRIX, L. W. (2017). Bridging finite and super population causal inference. *J*. *Causal Inference* **5** Art. No. 20160027, 8. MR4328880 https://doi.org/10.1515/jci-2016-0027

EGAMI, N., JACOBS-HARUKAWA, M., STEWART, B. M. and WEI, H. (2023). Using Large Language Model Annotations for Valid Downstream Statistical Inference in Social Science: Design-Based Semi-Supervised Learning. Preprint. Available at arXiv:2306.04746.

FESLER, L., DEE, T., BAKER, R. and EVANS, B. (2019). Text as data methods for education research. *J*. *Res*. *Educ*. *Eff*. **12** 707–727.

FOLTZ, P. W., LAHAM, D. and LANDAUER, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interact*. *Multimed*. *Electron*. *J*. *Comput*.-*Enhanc*. *Learn*. **1** 939–944.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann*. *Statist*. **19** 1–141. MR1091842 https://doi.org/10.1214/aos/1176347963

GEE, J. P. (2007). *Social Linguistics and Literacies*. Routledge, London.

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

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

JIANG, H. and NACHUM, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics* 702–712. PMLR.

JUZWIK, M. M., CURCIC, S., WOLBERS, K., MOXLEY, K. D., DIMLING, L. M. and SHANKLAND, R. K. (2006). Writing into the 21st century: An overview of research on writing, 1999 to 2004. *Writ*. *Commun*. **23** 451–476.

KIM, J. S., BURKHAUSER, M. A., MESITE, L. M., ASHER, C. A., RELYEA, J. E., FITZGERALD, J. and ELMORE, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *J*. *Educ*. *Psychol*. **113** 3.

KIM, Y., MOZER, R., MIRATIX, L. W. and AL-ADEIMI, S. (2024). ChatGPT vs. Machine learning: Assessing the efficacy and accuracy of large language models for automated essay scoring. Working paper.

KIZILCEC, R. F. and LEE, H. (2020). Algorithmic fairness in education. Preprint. Available at arXiv:2007.05443.

KUHN, M. et al. (2008). Building predictive models in R using the caret package. *J*. *Stat*. *Softw*. **28** 1–26.

KUHN, M. and JOHNSON, K. (2013). *Applied Predictive Modeling*. Springer, New York. MR3099297 https://doi.org/10.1007/978-1-4614-6849-3

LITMAN, D., ZHANG, H., CORRENTI, R., MATSUMURA, L. C. and WANG, E. (2021). A fairness evaluation of automated methods for scoring text evidence usage in writing. In *International Conference on Artificial Intelligence in Education* 255–267. Springer, Berlin.

LUCAS, C., NIELSEN, R. A., ROBERTS, M. E., STEWART, B. M., STORER, A. and TINGLEY, D. (2015). Computer-assisted text analysis for comparative politics. *Polit*. *Anal*. **23** 254–277.

MIRATRIX, L. W. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **75** 369–396. MR3021392 https://doi.org/10.1111/j.1467-9868.2012.01048.x

MOZER, R., MIRATRIX, L., RELYEA, J. E. and KIM, J. S. (2024). Combining human and automated scoring methods in experimental assessments of writing: A case study tutorial. *J. Educ. Behav. Stat.* **49** 780-816. https://doi.org/10.3102/10769986231207886

MOZER, R. and MIRATRIX, L. (2025). Supplement to "More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials." https://doi.org/10.1214/24-AOAS1967SUPPA, https://doi.org/10.1214/24-AOAS1967SUPPB

NEELAKANTAN, A., XU, T., PURI, R., RADFORD, A., HAN, J. M., TWOREK, J., YUAN, Q., TEZAK, N., KIM, J. W. et al. (2022). Text and code embeddings by contrastive pre-training. Preprint. Available at arXiv:2201.10005.

PASHLEY, N. E., BASSE, G. W. and MIRATRIX, L. W. (2021). Conditional as-if analyses in randomized experiments. *J. Causal Inference* **9** 264–284. MR4348656 https://doi.org/10.1515/jci-2021-0012

PENNEBAKER, J. W., FRANCIS, M. E. and BOOTH, R. J. (2001). *Linguistic Inquiry and Word Count*: *LIWC* 2001. *Mahway*: *Lawrence Erlbaum Associates* **71**.

RIPLEY, B. D. (2007). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. MR2451352

ROMERO, C. and VENTURA, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **33** 135–146.

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

SALDAÑA, J. (2021). The coding manual for qualitative researchers.

SALES, A. C., GAGNON-BARTSCH, J. A., BOTELHO, A., HEFFERNAN, N. T., WU, E. and MIRATRIX, L. W. (2019). ReLOOP: Precise Unbiased Estimation in Randomized Experiments Using Observational Auxilliary Data. Joint Statistical Meeting. Joint Statistical Meeting.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York. MR1140409 https://doi.org/10.1007/978-1-4612-4378-6

SHERMIS, M. D. and HAMNER, B. (2013). Contrasting state-of-the-art automated scoring of essays. In *Handbook of Automated Essay Evaluation* 313–346. Routledge, London.

SPLAWA-NEYMAN, J. DABROWSKA, D. M. and SPEED, T. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472.

TOULMIN, S. (2003). *The Uses of Argument*. Cambridge Univ. Press, Cambridge.

WANG, S., MCCORMICK, T. H. and LEEK, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proc. Natl. Acad. Sci. USA* **117** 30266–30275. MR4263300 https://doi.org/10.1073/pnas.2001238117

WEEGAR, R. and IDESTAM-ALMQUIST, P. (2023). Reducing workload in short answer grading using machine learning. *Int. J. Artif. Intell. Educ.* 1–27.

WOODS, B., ADAMSON, D., MIEL, S. and MAYFIELD, E. (2017). Formative essay feedback using predictive scoring models. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2071–2080.

WU, E. and GAGNON-BARTSCH, J. A. (2018). The LOOP estimator: Adjusting for covariates in randomized experiments. *Eval. Rev.* **42** 458–488. https://doi.org/10.1177/0193841x18808003

YAN, D., RUPP, A. A. and FOLTZ, P. W. (2020). *Handbook of Automated Scoring*: *Theory into Practice*. CRC Press, Boca Raton.

ZUPANC, K. and BOSNIC, Z. (2016). Advances in the field of automated essay evaluation. *Informatica* **39** 383–395.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 465–484
https://doi.org/10.1214/24-AOAS1968

TRACING THE IMPACTS OF MOUNT PINATUBO ERUPTION ON REGIONAL CLIMATE USING SPATIALLY-VARYING CHANGEPOINT DETECTION

BY SAMANTHA SHI-JUN[1,a], LYNDSAY SHAND[1,2,c] AND BO LI[1,b]

[1]*Department of Statistics, University of Illinois Urbana–Champaign,* [a]*wooimj2@illinois.edu,* [b]*libo@illinois.edu*
[2]*Statistical Sciences Department, Sandia National Laboratories,* [c]*lshand@sandia.gov*

Significant events, such as volcanic eruptions, can have global and long-lasting impacts on climate. These global impacts, however, are not uniform across space and time. Understanding how the Mt. Pinatubo eruption affects global and regional climate is of great interest for predicting the impact on climate due to similar events as well as understanding the possible effect of the stratospheric aerosol injections proposed to combat climate change. While many studies illustrated the impact of the Pinatubo eruption on a global scale, studies at a fine regional scale are scarce. We propose a Bayesian spatially-varying changepoint detection and estimation method to trace the impact of Mt. Pinatubo eruption on regional climate. Our approach takes into account the diffusing nature and spatial correlation of the climate changes attributed to the volcanic eruption. We illustrate our method and demonstrate its advantages over an existing changepoint detection method through simulations. Finally, we apply our method to monthly stratospheric aerosol optical depth and surface temperature data from 1985 to 1995 to detect and estimate changepoints following the 1991 Mt. Pinatubo eruption. Our results quantitatively characterize the spatial pattern of the eruption's impact on regional climate, complementing the previous studies on the global impact of the Pinatubo eruption.

REFERENCES

AMINIKHANGHAHI, S. and COOK, D. J. (2017). A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **51** 339–367. https://doi.org/10.1007/s10115-016-0987-z

BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. MR1150343 https://doi.org/10.1214/aos/1176348521

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

CLEVELAND, R. (1990). STL: A seasonal-trend decomposition procedure based on loess. *J. Off. Stat.* **6** 3–73.

DUTTON, E. G. and CHRISTY, J. R. (1992). Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichón and Pinatubo. *Geophys. Res. Lett.* **19** 2313–2316. https://doi.org/10.1029/92GL02495

GAO, C., ROBOCK, A. and AMMANN, C. (2008). Volcanic forcing of climate over the past 1500 years: An improved ice core-based index for climate models. *J. Geophys. Res., Atmos.* **113**.

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

GLOBAL MODELING AND ASSIMILATION OFFICE (GMAO) (2015b). MERRA-2 tavg1_2d_slv_Nx: 2d, 1-Hourly, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/VJAFPLI1CSIV

GLOBAL MODELING AND ASSIMILATION OFFICE (GMAO) (2015a). MERRA-2 tavg1_2d_aer_Nx: 2d, 1-Hourly, Time-averaged, Single-Level, Assimilation, Aerosol Diagnostics V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed June 15, 2022. https://doi.org/10.5067/KLICLTZ8EM9D

Key words and phrases. Aerosol optical depth, Bayesian hierarchical model, spatially-varying changepoint, temperature, volcanic eruption.

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475 https://doi.org/10.1198/016214502760047113

GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554 https://doi.org/10.3150/12-BEJSP06

GUINNESS, J. and FUENTES, M. (2016). Isotropic covariance functions on spheres: Some properties and modeling considerations. *J. Multivariate Anal.* **143** 143–152. MR3431424 https://doi.org/10.1016/j.jmva.2015.08.018

HALLEMA, D. W., SUN, G., CALDWELL, P. V., NORMAN, S. P., COHEN, E. C., LIU, Y., WARD, E. J. and MCNULTY, S. G. (2017). Assessment of wildland fire impacts on watershed annual water yield: Analytical framework and case studies in the United States. *Ecohydrology* **10** e1794.

HUANG, C., ZHANG, H. and ROBESON, S. M. (2011). On the validity of commonly used covariance and variogram functions on the sphere. *Math. Geosci.* **43** 721–733. MR2824128 https://doi.org/10.1007/s11004-011-9344-7

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 https://doi.org/10.1080/01621459.2012.737745

LABITZKE, K. and MCCORMICK, M. P. (1992). Stratospheric temperature increases due to Pinatubo aerosols. *Geophys. Res. Lett.* **19** 207–210. https://doi.org/10.1029/91GL02940

LENSKY, I. M., DAYAN, U. and HELMAN, D. (2018). Synoptic circulation impact on the near-surface temperature difference outweighs that of the seasonal signal in the eastern Mediterranean. *J. Geophys. Res., Atmos.* **123** 11–333.

MAJUMDAR, A., GELFAND, A. E. and BANERJEE, S. (2005). Spatio-temporal change-point modeling. *J. Statist. Plann. Inference* **130** 149–166. MR2128002 https://doi.org/10.1016/j.jspi.2003.08.022

MCCLAIN, C. R., SIGNORINI, S. R. and CHRISTIAN, J. R. (2004). Subtropical gyre variability observed by ocean-color satellites. *Deep-Sea Res., Part 2, Top. Stud. Oceanogr.* **51** 281–301.

MCCORMICK, M. and VEIGA, R. (1992). SAGE II measurements of early Pinatubo aerosols. *Geophys. Res. Lett.* **19** 155–158.

REEVES, J., CHEN, J., WANG, X. L., LUND, R. and LU, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteorol. Climatol.* **46** 900–915.

ROBBINS, M. W., LUND, R. B., GALLAGHER, C. M. and LU, Q. (2011). Changepoints in the North Atlantic tropical cyclone record. *J. Amer. Statist. Assoc.* **106** 89–99. MR2816704 https://doi.org/10.1198/jasa.2011.ap10023

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16** 351–367. MR1888450 https://doi.org/10.1214/ss/1015346320

ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. MR2749836 https://doi.org/10.1198/jcgs.2009.06134

ROBOCK, A. (2014). Stratospheric aerosol geoengineering. *Issues Environ. Sci. Technol.* **2014** 162–185. https://doi.org/10.1063/1.4916181

ROBOCK, A. and MAO, J. (1995). The volcanic signal in surface temperature observations. *J. Climate* **8** 1086–1103.

ROBOCK, A. and MATSON, M. (1983). Circumglobal transport of the el chichon volcanic dust cloud. *Science* **221** 195–197. https://doi.org/10.1126/science.221.4606.195

RYOO, J. M., WAUGH, D. W. and GETTELMAN, A. (2008). Variability of subtropical upper tropospheric humidity. *Atmos. Chem. Phys.* **8** 2643–2655.

SELF, S., ZHAO, J.-X., HOLASEK, R. E., TORRES, R. C. and KING, A. J. (1996). *The Atmospheric Impact of the* 1991 *Mount Pinatubo Eruption*. 1089–1115. Univ. Washington Press, Seattle.

SHAND, L. and LI, B. (2017). Modeling nonstationarity in space and time. *Biometrics* **73** 759–768. MR3713110 https://doi.org/10.1111/biom.12656

SHI-JUN, S., SHAND, L. and LI, B. (2025). Supplement to "Tracing the impacts of Mount Pinatubo eruption on regional climate using spatially-varying changepoint detection." https://doi.org/10.1214/24-AOAS1968SUPP

STEGLE, O., LIPPERT, C., MOOIJ, J., LAWRENCE, N. and BORGWARDT, K. (2011). Efficient inference in matrix-variate Gaussian models with iid observation noise. *Adv. Neural Inf. Process. Syst.* **24**.

STENCHIKOV, G., DELWORTH, T. L., RAMASWAMY, V., STOUFFER, R. J., WITTENBERG, A. and ZENG, F. (2009). Volcanic signals in oceans. *J. Geophys. Res., Atmos.* **114**. https://doi.org/10.1029/2008JD011673

STENCHIKOV, G., ROBOCK, A., RAMASWAMY, V., SCHWARZKOPF, M. D., HAMILTON, K. and RAMACHANDRAN, S. (2002). Arctic oscillation response to the 1991 Mount Pinatubo eruption: Effects of volcanic aerosols and ozone depletion. *J. Geophys. Res., Atmos.* **107** ACL-28.

STOWE, L., CAREY, R. and PELLEGRINO, P. (1992). Monitoring the Mt. Pinatubo aerosol layer with NOAA/11 AVHRR data. *Geophys. Res. Lett.* **19** 159–162.

THOMPSON, D. W. J., WALLACE, J. M., JONES, P. D. and KENNEDY, J. J. (2009). Identifying signatures of natural climate variability in time series of global-mean surface temperature: Methodology and insights. *J. Climate* **22** 6120–6141. https://doi.org/10.1175/2009JCLI3089.1

TUCKER, J. D. and YARGER, D. (2024). Elastic functional changepoint detection of climate impacts from localized sources. *Environmetrics* **35** Paper No. e2826, 18. MR4690073

VAN DYK, D. A. and JIAO, X. (2015). Metropolis-Hastings within partially collapsed Gibbs samplers. *J. Comput. Graph. Statist.* **24** 301–327. MR3357383 https://doi.org/10.1080/10618600.2014.930041

VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Amer. Statist. Assoc.* **103** 790–796. MR2524010 https://doi.org/10.1198/016214508000000409

WANG, M., HARRIS, T. and LI, B. (2023). Asynchronous changepoint estimation for spatially correlated functional time series. *J. Agric. Biol. Environ. Stat.* **28** 157–176. MR4546159 https://doi.org/10.1007/s13253-022-00519-w

XUAN, X. and MURPHY, K. (2007). Modeling changing dependency structure in multivariate time series. In *Proceedings of the* 24*th International Conference on Machine Learning* 1055–1062.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 485–504
https://doi.org/10.1214/24-AOAS1969

A SPATIAL AUTOREGRESSIVE RANDOM FOREST ALGORITHM FOR SMALL-AREA SPATIAL PREDICTION

BY CARA MACBRIDE[a], VINNY DAVIES[b] AND DUNCAN LEE[c]

School of Mathematics and Statistics, University of Glasgow, [a]*2392494m@student.gla.ac.uk,* [b]*vinny.davies@glasgow.ac.uk,* [c]*Duncan.Lee@glasgow.ac.uk*

In spatial areal unit data with missing or suppressed values, it is desirable to create models that are able to predict observations that are not available. Typically, statistical spatial smoothing models fitted in a Bayesian hierarchical framework are used for this purpose, which capture any unexplained residual spatial autocorrelation in the data through conditional autoregressive (CAR) or spatial autoregressive (SAR) priors applied to a set of random effects. In contrast, typical machine learning approaches, such as random forests or neural networks, ignore this residual autocorrelation and instead base predictions on complex nonlinear feature-target relationships. In this paper we propose `SPAR-Forest`, a novel spatial prediction algorithm that fuses random forests with spatial smoothing models. By iteratively refitting a random forest combined with a Bayesian CAR or SAR model in one algorithm, SPAR-Forest can incorporate flexible feature-target relationships while still accounting for the residual spatial autocorrelation. Our results, based on a Scottish property price data set and multiple simulated data sets, show that SPAR-Forest outperforms Bayesian CAR/SAR models, random forests, and state-of-the-art hybrid approaches, including geographical random forests, providing a state-of-the-art framework for small-area spatial prediction.

REFERENCES

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

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann*. *Inst*. *Statist*. *Math*. **43** 1–59. MR1105822 https://doi.org/10.1007/BF00116466

BOEHMKE, B. and GREENWELL, B. (2020). *Hands-on Machine Learning with R*. CRC Press/CRC, Boca Raton.

BREIMAN, L. (1984). *Classification and Regression Trees*. Routledge, London.

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

BREWER, M. J. and NOLAN, A. J. (2007). Variable smoothing in Bayesian intrinsic autoregressions. *Environmetrics* **18** 841–857. MR2416406 https://doi.org/10.1002/env.844

CREDIT, K. (2022). Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in Los Angeles. *Geogr. Anal*. **54** 58–83.

FARAWAY, J. J. (2015). *Linear Models with R*. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3186258

FRANKE, J. and NEUMANN, M. H. (2000). Bootstrapping neural networks. *Neural Comput*. **12** 1929–1949. https://doi.org/10.1162/089976600300015204

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

GEORGANOS, S., GRIPPA, T., GADIAGA, A., LINARD, C., LENNERT, M., VANHUYSSE, S., MBOGA, N., WOLFF, E. and KALOGIROU, S. (2021). Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Iinternational* **36** 121–136.

GREENWELL, B. and BOEHMKE, B. (2020). Variable importance plots—an introduction to the vip package. *R J*. **20** 343–366.

Key words and phrases. Areal unit data, random forests, property prices, spatial autoregressive models.

HAINING, B. and LI, G. (2020). *Modelling Spatial and Spatio-Temporal Data*: *A Bayesian Approach*. CRC Press/CRC, Boca Raton.

HENGL, T., HEUVELINK, G., KEMPEN, B., LEENAARS, J., WALSH, M., SHEPHERD, K., SILA, A., MACMILLAN, R., MENDES DE JESUS, J. et al. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* **10** 1–26.

JACK, E., LEE, D. and DEAN, N. (2019). Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model. *J. Roy. Statist. Soc. Ser. A* **182** 1061–1080. MR3955510 https://doi.org/10.1111/rssa.12447

KAWABATA, M., NAOI, M. and YASUDA, S. (2022). Earthquake risk reduction and residential land prices in Tokyo. *J. Spat. Econ.* **3** 5.

KIPF, T. and WELLING, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

KNORR-HELD, L. and RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.

KUHN, M. and JOHNSON, K. (2019). *Feature Engineering and Selection*: *A Practical Approach for Predictive Models*. CRC Press/CRC, Boca Raton.

LAWSON, A. B. (2009). *Bayesian Disease Mapping*: *Hierarchical Modeling in Spatial Epidemiology*. *Interdisciplinary Statistics*. CRC Press, Boca Raton, FL. MR2484272

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444. https://doi.org/10.1038/nature14539

LECUN, Y., BOSER, B., DENKER, J., HENDERSON, D., HOWARD, R., HUBBARD, W. and JACKEL, L. (1990). Handwritten digit recognition with a back-propagation network. *Adv. Neural Inf. Process. Syst.* 396–404.

LEE, D. and ANDERSON, C. (2023). Delivering spatially comparable inference on the risks of multiple severities of respiratory disease from spatially misaligned disease count data. *Biometrics* **79** 2691–2704. MR4644025 https://doi.org/10.1111/biom.13739

LEE, D., MEEKS, K. and PETTERSSON, W. (2021). Improved inference for areal unit count data using graph-based optimisation. *Stat. Comput.* **31** Paper No. 51. MR4280524 https://doi.org/10.1007/s11222-021-10025-7

LEROUX, B. G., LEI, X. and BRESLOW, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* (*Minneapolis*, *MN*, 1997) (M. Halloran and D. Berry, eds.). *IMA Vol. Math. Appl.* **116** 179–191. Springer, New York. MR1731684 https://doi.org/10.1007/978-1-4612-1284-3_4

MACBRIDE, C., DAVIES, V. and LEE, D. (2025a). Supplementary results and description to "A spatial autoregressive random forest algorithm for small-area spatial prediction." https://doi.org/10.1214/24-AOAS1969SUPPA

MACBRIDE, C., DAVIES, V. and LEE, D. (2025b). Supplementary software for "A spatial autoregressive random forest algorithm for small-area spatial prediction." https://doi.org/10.1214/24-AOAS1969SUPPB

PALMER, G., DU, S., POLITOWICZ, A., EMORY, J., YANG, X., GAUTAM, A., GUPTA, G., LI, Z., JACOBS, R. et al. (2022). Calibration after bootstrap for accurate uncertainty quantification in regression models. *Npj Comput. Mater.* **8** 115.

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

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

SAHA, A., BASU, S. and DATTA, A. (2023). Random forests for spatially dependent data. *J. Amer. Statist. Assoc.* **118** 665–683. MR4571149 https://doi.org/10.1080/01621459.2021.1950003

SOLTANI, A., HEYDARI, M., AGHAEI, F. and PETTIT, C. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities* **131** 103941.

WANG, X., WANG, X., REN, X. and WEN, F. (2022). Can digital financial inclusion affect CO2 emissions of China at the prefecture level? Evidence from a spatial econometric approach. *Energy Econ.* **109** 105966.

WHITTLE, P. (1954). On stationary processes in the plane. *Biometrika* **41** 434–449. MR0067450 https://doi.org/10.1093/biomet/41.3-4.434

WRIGHT, M. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77** 1–17.

XIA, Z., STEWART, K. and FAN, J. (2021). Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major U.S. metropolitan area. *Comput. Environ. Urban Syst.* **87** 101599.

ZHANG, H., ZIMMERMAN, J., NETTLETON, D. and NORDMAN, D. J. (2020). Random forest prediction intervals. *Amer. Statist.* **74** 392–406. MR4168259 https://doi.org/10.1080/00031305.2019.1585288

ZHU, D., LIU, Y., YAO, X. and FISCHER, M. (2022). Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions. *Geoinformatica* **26** 645–676.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 505–528
https://doi.org/10.1214/24-AOAS1970

DYNAMIC PREDICTION WITH MULTIVARIATE LONGITUDINAL OUTCOMES AND LONGITUDINAL MAGNETIC RESONANCE IMAGING DATA

BY HAOTIAN ZOU[1,a], LUO XIAO[2,c], DONGLIN ZENG[3,d] AND SHENG LUO[1,b]

For the Alzheimer's Disease Neuroimaging Initiative[*]

[1]*Department of Biostatistics and Bioinformatics, Duke University,* [a]*haotian.zou@duke.edu,* [b]*sheng.luo@duke.edu*
[2]*Department of Statistics, North Carolina State University,* [c]*lxiao5@ncsu.edu*
[3]*Department of Biostatistics, University of Michigan,* [d]*dzeng@email.unc.edu*

Alzheimer's Disease (AD) is a common neurodegenerative disorder impairing multiple domains. Recent AD studies, for example, the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, collect multimodal data to better understand AD severity and progression. To facilitate precision medicine for high-risk individuals, it is essential to develop an AD predictive model that leverages multimodal data and provides accurate personalized predictions of dementia occurrences. In this article we propose a multivariate functional mixed model with longitudinal magnetic resonance imaging data (MFMM-LMRI) that jointly models longitudinal neurological scores, longitudinal voxelwise MRI data, and the survival outcome as dementia onset. We model longitudinal MRI data using the joint and individual variation explained (JIVE) approach. We investigate two functional forms linking the longitudinal and survival processes. We adopt the Markov chain Monte Carlo (MCMC) method to obtain posterior samples. We establish a dynamic prediction framework that predicts longitudinal trajectories and the probability of dementia occurrence. The simulation study with various sample sizes and event rates supports the validity of the method. We apply the MFMM-LMRI to the motivating ADNI study and conclude that additional ApoE-ϵ4 alleles and a higher latent disease profile are associated with a higher risk of dementia onset. We detect a significant association between the longitudinal MRI data and the survival outcome. The instantaneous model with longitudinal MRI data has the best fitting and predictive performance.

REFERENCES

ALBERT, M. S., DEKOSKY, S. T., DICKSON, D., DUBOIS, B., FELDMAN, H. H., FOX, N. C., GAMST, A., HOLTZMAN, D. M., JAGUST, W. J. et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement*. **7** 270–279.

ALEXANDER, A. L., LEE, J. E., LAZAR, M. and FIELD, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics* **4** 316–329.

ANDREWS, S. J., FULTON-HOWARD, B. and GOATE, A. (2020). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol*. **19** 326–335.

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 multiomics data sets. *Mol. Syst. Biol*. **14** e8124. https://doi.org/10.15252/msb.20178124

ASHBURNER, J. and FRISTON, K. J. (2000). Voxel-based morphometry–the methods. *NeuroImage* **11** 805–821. https://doi.org/10.1006/nimg.2000.0582

BAILEY, D. L., MAISEY, M. N., TOWNSEND, D. W. and VALK, P. E. (2005). *Positron Emission Tomography* **2**. Springer, Berlin.

Key words and phrases. Alzheimer's Disease, functional data, longitudinal magnetic resonance imaging data, joint model.

Busatto, G. F., Garrido, G. E., Almeida, O. P., Castro, C. C., Camargo, C. H., Cid, C. G., Buchpiguel, C. A., Furuie, S. and Bottino, C. M. (2003). A voxel-based morphometry study of temporal lobe gray matter reductions in Alzheimer's disease. *Neurobiol. Aging* **24** 221–231.

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

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

Chouraki, V., Reitz, C., Maury, F., Bis, J. C., Bellenguez, C., Yu, L., Jakobsdottir, J., Mukherjee, S., Adams, H. H. et al. (2016). Evaluation of a genetic risk score to improve risk prediction for Alzheimer's disease. *J. Alzheimer's Dis.* **53** 921–932.

Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G. W., Roses, A. D., Haines, J. L. and Pericak-Vance, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261** 921–923. https://doi.org/10.1126/science.8346443

de Wilde, A., van Maurik, I. S., Kunneman, M., Bouwman, F., Zwan, M., Willemse, E. A., Biessels, G. J., Minkman, M., Pel, R. et al. (2017). Alzheimer's biomarkers in daily practice (ABIDE) project: Rationale and design. *Alzheimer's Dement.* **6** 143–151.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Stat.* **3** 458–488. MR2668715 https://doi.org/10.1214/08-AOAS206

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., Badarinarayan, N., Gerad/Perades, IGAP consortia et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* **138** 3673–3684.

Feng, S., Wolfe, R. A. and Port, F. K. (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using Poisson variance structures. *J. Amer. Statist. Assoc.* **100** 728–735. MR2206989 https://doi.org/10.1198/016214505000000123

Galimard, J.-E., Chevret, S., Protopopescu, C. and Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat. Med.* **35** 2907–2920. MR3528233 https://doi.org/10.1002/sim.6902

Goerdten, J., Čukić, I., Danso, S. O., Carrière, I. and Muniz-Terrera, G. (2019). Statistical methods for dementia risk prediction and recommendations for future work: A systematic review. *Alzheimer's Dement.* **5** 563–569.

Goldsmith, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 453–469. MR2914521 https://doi.org/10.1111/j.1467-9876.2011.01031.x

Goldsmith, J. and Kitago, T. (2016). Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 215–236. MR3456686 https://doi.org/10.1111/rssc.12115

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18** 2529–2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5

Greven, S., Crainiceanu, C., Caffo, B. and Reich, D. (2011). Longitudinal functional principal component analysis. In *Recent Advances in Functional Data Analysis and Related Topics*. *Contrib. Statist.* 149–154. Physica-Verlag/Springer, Heidelberg. MR2815575 https://doi.org/10.1007/978-3-7908-2736-1_23

Gross, A. L., Mungas, D. M., Leoutsakos, J.-M. S., Albert, M. S., Jones, R. N., Alzheimer's Disease Neuroimaging Initiative et al. (2016). Alzheimer's disease severity, objectively determined and measured. *Alzheimer's Dement.* **4** 159–168.

Hammon, A. and Zinn, S. (2020). Multiple imputation of binary multilevel missing not at random data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 547–564. MR4098961 https://doi.org/10.1111/rssc.12401

Hasenstab, K., Scheffler, A., Telesca, D., Sugar, C. A., Jeste, S., DiStefano, C. and Şentürk, D. (2017). A multi-dimensional functional principal components analysis of EEG data. *Biometrics* **73** 999–1009. MR3713133 https://doi.org/10.1111/biom.12635

Hasin, Y., Seldin, M. and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* **18** 1–15.

Hickey, G. L., Philipson, P., Jorgensen, A. and Kolamunnage-Dona, R. (2018). Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. *Int. J. Biostat.* **14** 20170047. MR3812146 https://doi.org/10.1515/ijb-2017-0047

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

Huang, M., Yang, W., Feng, Q. and Chen, W. (2017). Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. *Sci. Rep.* **7** 1–13.

HUGHES, C. P., BERG, L., DANZIGER, W., COBEN, L. A. and MARTIN, R. L. (1982). A new clinical scale for the staging of dementia. *Br. J. Psychiatry* **140** 566–572.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 https://doi.org/10.1214/08-AOAS169

JIANG, S., CAO, J., ROSNER, B. and COLDITZ, G. A. (2023). Supervised two-dimensional functional principal component analysis with time-to-event outcomes and mammogram imaging data. *Biometrics* **79** 1359–1369. MR4606358 https://doi.org/10.1111/biom.13611

JIANG, S., XIE, Y. and COLDITZ, G. A. (2021). Functional ensemble survival tree: Dynamic prediction of Alzheimer's disease progression accommodating multiple time-varying covariates. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 66–79. MR4204938 https://doi.org/10.1111/rssc.12449

KANG, K. and SONG, X. Y. (2023). Joint modeling of longitudinal imaging and survival data. *J. Comput. Graph. Statist.* **32** 402–412. MR4592919 https://doi.org/10.1080/10618600.2022.2102027

KATINA, S., VITTERT, L. and BOWMAN, A. W. (2021). Functional data analysis and visualisation of three-dimensional surface shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 691–713. MR4275842 https://doi.org/10.1111/rssc.12482

KATZMAN, J. L., SHAHAM, U., CLONINGER, A., BATES, J., JIANG, T. and KLUGER, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18** 1–12.

KEIHANINEJAD, S., ZHANG, H., RYAN, N. S., MALONE, I. B., MODAT, M., CARDOSO, M. J., CASH, D. M., FOX, N. C. and OURSELIN, S. (2013). An unbiased longitudinal analysis framework for tracking white matter changes using diffusion tensor imaging with application to Alzheimer's disease. *NeuroImage* **72** 153–163.

KNOPMAN, D. S., AMIEVA, H., PETERSEN, R. C., CHÉTELAT, G., HOLTZMAN, D. M., HYMAN, B. T., NIXON, R. A. and JONES, D. T. (2021). Alzheimer disease. *Nat. Rev. Dis. Primers* **7** 1–21.

KONG, D., GIOVANELLO, K. S., WANG, Y., LIN, W., LEE, E., FAN, Y., MURALI DORAISWAMY, P., ZHU, H., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2015). Predicting Alzheimer's disease using combined imaging-whole genome SNP data. *J. Alzheimer's Dis.* **46** 695–702.

KONG, D., IBRAHIM, J. G., LEE, E. and ZHU, H. (2018). FLCRM: Functional linear Cox regression model. *Biometrics* **74** 109–117. MR3777931 https://doi.org/10.1111/biom.12748

LAWLESS, J. and ZHAN, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canad. J. Statist.* **26** 549–565.

LEUZY, A., SMITH, R., CULLEN, N. C., STRANDBERG, O., VOGEL, J. W., BINETTE, A. P., BORRONI, E., JANELIDZE, S., OHLSSON, T. et al. (2022). Biomarker-based prediction of longitudinal Tau positron emission tomography in Alzheimer disease. *JAMA Neurol.* **79** 149–158.

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

LI, C., XIAO, L. and LUO, S. (2022). Joint model for survival and multivariate sparse functional data with application to a study of Alzheimer's Disease. *Biometrics* **78** 435–447. MR4450566 https://doi.org/10.1111/biom.13427

LI, H., HABES, M., WOLK, D. A., FAN, Y., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2019). A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer's Dement.* **15** 1059–1070.

LI, K., CHAN, W., DOODY, R. S., QUINN, J., LUO, S. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2017). Prediction of conversion to Alzheimer's Disease with longitudinal measures and time-to-event data. *J. Alzheimer's Dis.* **58** 361–371. https://doi.org/10.3233/JAD-161201

LI, K. and LUO, S. (2019). Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease. *Stat. Methods Med. Res.* **28** 327–342. MR3903744 https://doi.org/10.1177/0962280217722177

LI, K., O'BRIEN, R., LUTZ, M., LUO, S., INITIATIVE, A. D. N. et al. (2018). A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data. *Alzheimer's Dement.* **14** 644–651.

LI, L., GREENE, T. and HU, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat. Methods Med. Res.* **27** 2264–2278. MR3825906 https://doi.org/10.1177/0962280216680239

LI, R. and XIAO, L. (2023). Latent factor model for multivariate functional data. *Biometrics* **79** 3307–3318. MR4680723 https://doi.org/10.1111/biom.13924

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

LONG, J. D. and MILLS, J. A. (2018). Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Med. Res. Methodol.* **18** 1–15.

MAUFF, K., STEYERBERG, E., KARDYS, I., BOERSMA, E. and RIZOPOULOS, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. *Stat*. *Comput*. **30** 999–1014. MR4108688 https://doi.org/10.1007/s11222-020-09927-9

MILÀ-ALOMÀ, M., SALVADÓ, G., GISPERT, J. D., VILOR-TEJEDOR, N., GRAU-RIVERA, O., SALA-VILA, A., SÁNCHEZ-BENAVIDES, G., ARENAZA-URQUIJO, E. M., CROUS-BOU, M. et al. (2020). Amyloid beta, Tau, synaptic, neurodegeneration, and glial biomarkers in the preclinical stage of the Alzheimer's continuum. *Alzheimer's Dement*. **16** 1358–1371.

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

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

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu*. *Rev*. *Stat*. *Appl*. **6** 223–240. MR3939519 https://doi.org/10.1146/annurev-statistics-030718-105048

PARK, S. Y. and STAICU, A.-M. (2015). Longitudinal functional data analysis. *Stat* **4** 212–226. MR3405402 https://doi.org/10.1002/sta4.89

PENNANEN, C., TESTA, C., LAAKSO, M., HALLIKAINEN, M., HELKALA, E., HÄNNINEN, T., KIVIPELTO, M., KÖNÖNEN, M., NISSINEN, A. et al. (2005). A voxel based morphometry study on mild cognitive impairment. *J*. *Neurol*. *Neurosurg*. *Psychiatry* **76** 11–14.

PETERSEN, R. C. (2016). Mild cognitive impairment. *Continuum* (*Minneap*. *Minn*.) **22** 404–418. https://doi.org/10.1212/CON.0000000000000313

POLZEHL, J. and TABELOW, K. (2019). *Magnetic Resonance Brain Imaging*. Springer, Berlin.

PROUST-LIMA, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach. *Stat*. *Med*. **35** 382–398. MR3455508 https://doi.org/10.1002/sim.6731

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat*. *Med*. **30** 1366–1380. MR2828959 https://doi.org/10.1002/sim.4205

ROSEN, W. G., MOHS, R. C. and DAVIS, K. L. (1984). A new rating scale for Alzheimer's disease. *Am*. *J*. *Psychiatr.* **141** 1356–1364. https://doi.org/10.1176/ajp.141.11.1356

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

SHI, H., MA, D., NIE, Y., BEG, M. F., PEI, J., CAO, J., ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE et al. (2021). Early diagnosis of Alzheimer's disease on ADNI data using novel longitudinal score based on functional principal component analysis. *J*. *Med*. *Imag*. **8** 024502–024502.

SONG, X., XIA, Y. and ZHU, H. (2017). Hidden Markov latent variable models with multivariate longitudinal data. *Biometrics* **73** 313–323. MR3632377 https://doi.org/10.1111/biom.12536

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

TANG, A.-M., TANG, N.-S. and ZHU, H. (2017). Influence analysis for skew-normal semiparametric joint models of multivariate longitudinal and multivariate survival data. *Stat*. *Med*. **36** 1476–1490. MR3631973 https://doi.org/10.1002/sim.7211

TOMBAUGH, T. N. and MCINTYRE, N. J. (1992). The mini-mental state examination: A comprehensive review. *J*. *Amer. Geriatr. Soc*. **40** 922–935.

VAN MAURIK, I. S., VOS, S. J., BOS, I., BOUWMAN, F. H., TEUNISSEN, C. E., SCHELTENS, P., BARKHOF, F., FROLICH, L., KORNHUBER, J. et al. (2019). Biomarker-based prognosis for people with mild cognitive impairment (ABIDE): A modelling study. *Lancet Neurol*. **18** 1034–1044.

VAN MAURIK, I. S., ZWAN, M. D., TIJMS, B. M., BOUWMAN, F. H., TEUNISSEN, C. E., SCHELTENS, P., WATTJES, M. P., BARKHOF, F., BERKHOF, J. et al. (2017). Interpreting biomarker results in individual patients with mild cognitive impairment in the Alzheimer's biomarkers in daily practice (ABIDE) project. *JAMA Neurol*. **74** 1481–1491.

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal*. **16** 667–718. MR4298989 https://doi.org/10.1214/20-ba1221

VEITCH, D. P., WEINER, M. W., AISEN, P. S., BECKETT, L. A., DECARLI, C., GREEN, R. C., HARVEY, D., JACK JR, C. R., JAGUST, W. et al. (2022). Using the Alzheimer's Disease Neuroimaging Initiative to improve early detection, diagnosis, and treatment of Alzheimer's disease. *Alzheimer's Dement*. **18** 824–857.

VLAARDINGERBROEK, M. T. and BOER, J. A. (2013). *Magnetic Resonance Imaging*: *Theory and Practice*. Springer, Berlin.

WANG, J., LUO, S. and LI, L. (2017). Dynamic prediction for multiple repeated measures and event time data: An application to Parkinson's disease. *Ann*. *Appl*. *Stat*. **11** 1787–1809. MR3709578 https://doi.org/10.1214/17-AOAS1059

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

XIAO, L., LI, Y. and RUPPERT, D. (2013). Fast bivariate P-splines: The sandwich smoother. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **75** 577–599. MR3065480 https://doi.org/10.1111/rssb.12007

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

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

ZHANG, M. J., XIA, F. and ZOU, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nat*. *Commun*. **10** 1–11.

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

ZHOU, H., ZHANG, X. and CHEN, J. (2021). Covariate adaptive familywise error rate control for genome-wide association studies. *Biometrika* **108** 915–931. MR4341359 https://doi.org/10.1093/biomet/asaa098

ZHU, H., MORRIS, J. S., WEI, F. and COX, D. D. (2017). Multivariate functional response regression, with application to fluorescence spectroscopy in a cervical pre-cancer study. *Comput*. *Statist*. *Data Anal*. **111** 88–101. MR3630220 https://doi.org/10.1016/j.csda.2017.02.004

ZIPUNNIKOV, V., CAFFO, B., YOUSEM, D. M., DAVATZIKOS, C., SCHWARTZ, B. S. and CRAINICEANU, C. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage* **58** 772–784. https://doi.org/10.1016/j.neuroimage.2011.05.085

ZOU, H., LI, K., ZENG, D., LUO, S. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2021). Bayesian inference and dynamic prediction of multivariate joint model with functional data: An application to Alzheimer's disease. *Stat*. *Med*. **40** 6855–6872. MR4352771 https://doi.org/10.1002/sim.9214

ZOU, H., XIAO, L., ZENG, D. and LUO, S. (2025a). Supplementary Material for "Dynamic Prediction with Multivariate Longitudinal Outcomes and Longitudinal Magnetic Resonance Imaging Data." https://doi.org/10.1214/24-AOAS1970SUPPA.

ZOU, H., XIAO, L., ZENG, D. and LUO, S. (2025b). Code for "Dynamic Prediction with Multivariate Longitudinal Outcomes and Longitudinal Magnetic Resonance Imaging Data." https://doi.org/10.1214/24-AOAS1970SUPPB.

ZOU, H., ZENG, D., XIAO, L. and LUO, S. (2023). Bayesian inference and dynamic prediction for multivariate longitudinal and survival data. *Ann*. *Appl*. *Stat*. **17** 2574–2595. MR4637681 https://doi.org/10.1214/23-aoas1733

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 529–565
https://doi.org/10.1214/24-AOAS1971

IDENTIFYING PEER INFLUENCE IN THERAPEUTIC COMMUNITIES ADJUSTING FOR LATENT HOMOPHILY

BY SHANJUKTA NATH[1,a], KEITH WARREN[2,b] AND SUBHADEEP PAUL[3,c]

[1]*Department of Agricultural and Applied Economics, University of Georgia,* [a]*shanjukta.nath@uga.edu*
[2]*College of Social Work, The Ohio State University,* [b]*warren.193@osu.edu*
[3]*Department of Statistics, The Ohio State University,* [c]*paul.963@osu.edu*

We investigate peer role model influence on successful graduation from therapeutic communities (TCs) for substance abuse and criminal behavior. We use data from three TCs that kept records of exchanges of affirmations among residents and their precise entry and exit dates, allowing us to form peer networks and define a causal effect of interest. The role model effect measures the difference in the expected outcome of a resident (ego) who can observe one of their peers graduate before the ego's exit vs. not graduating. To identify peer influence in the presence of unobserved homophily in observational data, we model the network with a latent variable model. We show that our peer influence estimator is asymptotically unbiased when the unobserved latent positions are estimated from the observed network. We additionally propose a measurement error bias correction method to further reduce bias due to estimating latent positions. Our simulations show the proposed latent homophily adjustment and bias correction perform well in finite samples. We also extend the methodology to the case of binary response with a probit model. Our results indicate a positive effect of peers' graduation on residents' graduation and that it differs based on gender, race, and the definition of the role model effect. A counterfactual exercise quantifies the potential benefits of an intervention directly on the treated resident and indirectly on their peers through network propagation.

REFERENCES

ANDREWS, D. A. and BONTA, J. (1995). *The Level of Service Inventory-Revised* [*User Manual*]. Multi-health systems, Toronto, ON.

ARAL, S. and NICOLAIDES, C. (2017). Exercise contagion in a global social network. *Nat. Commun.* **8** 14753.

ATHREYA, A., FISHKIND, D. E., TANG, M., PRIEBE, C. E., PARK, Y., VOGELSTEIN, J. T., LEVIN, K., LYZINSKI, V., QIN, Y. et al. (2017). Statistical inference on random dot product graphs: A survey. *J. Mach. Learn. Res.* **18** Paper No. 226, 92. MR3827114

ATHREYA, A., PRIEBE, C. E., TANG, M., LYZINSKI, V., MARCHETTE, D. J. and SUSSMAN, D. L. (2016). A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* **78** 1–18. MR3494576 https://doi.org/10.1007/s13171-015-0071-x

BAIRD, M. D., ENGBERG, J. and OPPER, I. M. (2023). Optimal allocation of seats in the presence of peer effects: Evidence from a job training program. *J. Labor Econ.* **41** 479–509.

BASSE, G., DING, P., FELLER, A. and TOULIS, P. (2024). Randomization tests for peer effects in group formation experiments. *Econometrica* **92** 567–590. MR4730816 https://doi.org/10.3982/ecta20134

BAYER, P., HJALMARSSON, R. and POZEN, D. (2009). Building criminal capital behind bars: Peer effects in juvenile corrections. *Q. J. Econ.* **124** 105–147.

BEST, D. (2019). What we know about recovery, desistance and reintegration. In *Pathways to Recovery and Desistance* 1–22. Policy Press.

BOSTWICK, V. K. and WEINBERG, B. A. (2022). Nevertheless she persisted? Gender peer effects in doctoral stem programs. *J. Labor Econ.* **40** 397–436.

BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993 https://doi.org/10.1016/j.jeconom.2008.12.021

Key words and phrases. Networks, peer influence, latent homophily, causal inference.

CAMPBELL, B., WARREN, K., WEILER, M. and DE LEON, G. (2021). Eigenvector centrality defines hierarchy and predicts graduation in therapeutic community units. *PLoS ONE* **16** e0261405.

CHAN, T. Y., LI, J. and PIERCE, L. (2014). Compensation and peer effects in competing sales teams. *Manage. Sci.* **60** 1965–1984.

CHRISTAKIS, N. A. and FOWLER, J. H. (2007). The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357** 370–379.

DART ROXBURGH, A., BEST, D., LUBMAN, D. I. and MANNING, V. (2023). Composition of social networks to build recovery capital differ across early and stable stages of recovery. *Addict. Res. Theory* 1–8.

DAVIES, R. B. and HUTTON, B. (1975). The effect of errors in the independent variables in linear regression. *Biometrika* **62** 383–391. MR0395065 https://doi.org/10.2307/2335377

DE LEON, G. (2000). *The Therapeutic Community: Theory, Model, and Method.*

DE LEON, G., WEXLER, H. K. and JAINCHILL, N. (1982). The therapeutic community: Success and improvement rates 5 years after treatment. *Int. J. Addict.* **17** 703–747.

DEAN, D. O., BAUER, D. J. and PRINSTEIN, M. J. (2017). Friendship dissolution within social networks modeled through multilevel event history analysis. *Multivar. Behav. Res.* **52** 271–289.

DEMING, D. J. (2011). Better schools, less crime? *Q. J. Econ.* **126** 2063–2115.

DENNING, J. T., MURPHY, R. and WEINHARDT, F. (2023). Class rank and long-run outcomes. *Rev. Econ. Stat.* 1–16.

DICICCIO, C. J. and ROMANO, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *J. Amer. Statist. Assoc.* **112** 1211–1220. MR3735371 https://doi.org/10.1080/01621459.2016.1202117

DOOGAN, N. J. and WARREN, K. (2017). A network of helping: Generalized reciprocity and cooperative behavior in response to peer and staff affirmations and corrections among therapeutic community residents. *Addict. Res. Theory* **25** 243–250.

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

GAVIRIA, A. and RAPHAEL, S. (2001). School-based peer effects and juvenile behavior. *Rev. Econ. Stat.* **83** 257–268.

GERSHENSON, S., MD HART, C., HYMAN, J., LINDSAY, C. A. and PAPAGEORGE, N. W. (2022). The long-run impacts of same-race teachers. *Amer. Econ. J.: Econ. Policy* **14** 300–342.

GOLDSMITH-PINKHAM, P. and IMBENS, G. W. (2013). Social networks and the identification of peer effects. *J. Bus. Econom. Statist.* **31** 253–264. MR3173674 https://doi.org/10.1080/07350015.2013.801251

GONG, J., LU, Y. and SONG, H. (2021). Gender peer effects on students' academic and noncognitive outcomes: Evidence and mechanisms. *J. Hum. Resour.* **56** 686–710.

GOSSOP, M. (2000). The therapeutic community: Theory, model and method. *Addiction* **95** 1720.

HERBST, D. and MAS, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science* **350** 545–549.

HOFF, P. (2021). Additive and multiplicative effects network models. *Statist. Sci.* **36** 34–50. MR4194202 https://doi.org/10.1214/19-STS757

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

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

HSIEH, C.-S. and LEE, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *J. Appl. Econometrics* **31** 301–319. MR3481364 https://doi.org/10.1002/jae.2426

JASON, L. A., LYNCH, G., BOBAK, T., LIGHT, J. M. and DOOGAN, N. J. (2022). Dynamic interdependence of advice seeking, loaning, and recovery characteristics in recovery homes. *J. Hum. Behav. Soc. Environ.* **32** 663–678.

KIESSLING, L. and NORRIS, J. (2023). The long-run effects of peers on mental health. *Econ. J.* **133** 281–322.

KREAGER, D. A., SCHAEFER, D. R., DAVIDSON, K. M., ZAJAC, G., HAYNIE, D. L. and DE LEON, G. (2019). Evaluating peer-influence processes in a prison-based therapeutic community: A dynamic network approach. *Drug Alcohol Depend.* **203** 13–18.

LAZER, D. (2001). The co-evolution of individual and network. *J. Math. Sociol.* **25** 69–108.

LAZER, D., RUBINEAU, B., CHETKOVICH, C., KATZ, N. and NEBLO, M. (2010). The coevolution of networks and political attitudes. *Polit. Commun.* **27** 248–274.

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

LI, T., LEVINA, E. and ZHU, J. (2020). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. MR4108931 https://doi.org/10.1093/biomet/asaa006

LI, X., DING, P., LIN, Q., YANG, D. and LIU, J. S. (2019). Randomization inference for peer effects. *J. Amer. Statist. Assoc.* **114** 1651–1664. MR4047290 https://doi.org/10.1080/01621459.2018.1512863

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

MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud*. **60** 531–542. MR1236836 https://doi.org/10.2307/2298123

MCCORKEL, J., HARRISON, L. D. and INCIARDI, J. A. (1998). How treatment is constructed among graduates and dropouts in a prison therapeutic community for women. *J. Offender Rehabil*. **27** 37–59.

MCFOWLAND, E. III and SHALIZI, C. R. (2023). Estimating causal peer influence in homophilous social networks by inferring latent locations. *J. Amer. Statist. Assoc*. **118** 707–718. MR4571152 https://doi.org/10.1080/01621459.2021.1953506

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

MERICLE, A. A., HOWELL, J., BORKMAN, T., SUBBARAMAN, M. S., FISHER SANDERS, B. and POLCIN, D. L. (2023). Social model recovery and recovery housing. *Addict. Res. Theory* 1–8.

NAKAMURA, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77** 127–137. MR1049414 https://doi.org/10.1093/biomet/77.1.127

NATH, S., WARREN, K. and PAUL, S. (2025). Supplement to "Identifying peer influence in therapeutic communities adjusting for latent homophily." https://doi.org/10.1214/24-AOAS1971SUPPA, https://doi.org/10.1214/24-AOAS1971SUPPB

NOVICK, S. J. and STEFANSKI, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *J. Amer. Statist. Assoc*. **97** 472–481. MR1941464 https://doi.org/10.1198/016214502760047005

PAUL, S. and CHEN, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Stat*. **10** 3807–3870. MR3579677 https://doi.org/10.1214/16-EJS1211

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

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

RIESSMAN, F. (1965). The "helper" therapy principle. *Soc. Netw.* 27–32.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist*. **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

RUBIN-DELANCHY, P., CAPE, J., TANG, M. and PRIEBE, C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **84** 1446–1473. MR4494166 https://doi.org/10.1111/rssb.12509

SACERDOTE, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education* **3** 249–277. Elsevier, Amsterdam.

SCHAFER, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika* **74** 385–391. MR0903139 https://doi.org/10.1093/biomet/74.2.385

SCHAFER, D. W. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrika* **80** 899–904.

SHALIZI, C. R. and THOMAS, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res*. **40** 211–239. MR2767833 https://doi.org/10.1177/0049124111404820

SNIJDERS, T. A. B. (2017). Stochastic actor-oriented models for network dynamics. *Annu. Rev. Stat. Appl*. **4** 343–363.

SRIDHAR, D., DE BACCO, C. and BLEI, D. (2022). Estimating social influence from observational data. In *Conference on Causal Learning and Reasoning* 712–733. PMLR.

STEFANSKI, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72** 583–592. MR0817572 https://doi.org/10.1093/biomet/72.3.583

STEFANSKI, L. A. and CARROLL, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist*. **13** 1335–1351. MR0811496 https://doi.org/10.1214/aos/1176349741

STEVENSON, M. (2017). Breaking bad: Mechanisms of social influence and the path to criminality in juvenile jails. *Rev. Econ. Stat*. **99** 824–838.

SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc*. **107** 1119–1128. MR3010899 https://doi.org/10.1080/01621459.2012.699795

TANG, M. and PRIEBE, C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist*. **46** 2360–2415. MR3845021 https://doi.org/10.1214/17-AOS1623

TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. In *International Conference on Machine Learning* 1489–1497. PMLR.

VANDERWEELE, T. J. (2011). Sensitivity analysis for contagion effects in social networks. *Sociol. Methods Res.* **40** 240–255. MR2767834 https://doi.org/10.1177/0049124111404821

VANDERWEELE, T. J. and AN, W. (2013). Social networks and causal inference. In *Handbook of Causal Analysis for Social Research* 353–374. Springer, Berlin.

WARREN, K., CAMPBELL, B. and CRANMER, S. (2020). Tightly bound: The relationship of network clustering coefficients and reincarceration at three therapeutic communities. *J. Stud. Alcohol Drugs* **81** 673–680.

WARREN, K., CAMPBELL, B., CRANMER, S., DE LEON, G., DOOGAN, N., WEILER, M. and DOHERTY, F. (2020). Building the community: Endogenous network formation, homophily and prosocial sorting among therapeutic community residents. *Drug Alcohol Depend.* **207** 107773.

WARREN, K., DOOGAN, N. J. and DOHERTY, F. (2021). Difference in response to feedback and gender in three therapeutic community units. *Front. Psychiatr.* **12** 690713.

WARREN, K. L. (2020). Senior therapeutic community members show greater consistency when affirming peers: Evidence of social learning. *Ther. Communities*.

WARREN, K. L., DOOGAN, N., WERNEKINCK, U. and DOHERTY, F. C. (2021). Resident interactions when affirming and correcting peers in a therapeutic community for women. *Ther. Communities*.

WHITE, W. L. and KURTZ, E. (2008). Twelve defining moments in the history of alcoholics anonymous. In *Recent Developments in Alcoholism*: *Research on Alcoholics Anonymous and Spirituality in Addiction Recovery* 37–57.

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

XIE, F. and XU, Y. (2023). Efficient estimation for random dot product graphs via a one-step procedure. *J. Amer. Statist. Assoc.* **118** 651–664. MR4571148 https://doi.org/10.1080/01621459.2021.1948419

YATES, R., BURNS, J. and MCCABE, L. (2017). Integration: Too much of a bad thing? *Journal of Groups in Addiction & Recovery* **12** 196–206.

ZIMMERMAN, S. D. (2019). Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review* **109** 1–47.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 566–585
https://doi.org/10.1214/24-AOAS1972

EBICOP: ENSEMBLE BIVARIATE COPULAS FOR MODELING MULTIVARIATE CYBER DATA BREACH RISKS

BY YIJIA LI[1,a], QUYNH NHU NGUYEN[2,c], MAOCHAO XU[3,d], PENG ZHAO[4,e] AND TAIZHONG HU[1,b]

[1]*Department of Statistics and Finance, University of Science and Technology of China,* [a]*liyijia@mail.ustc.edu.cn,* [b]*thu@ustc.edu.cn*

[2]*Department of Management Science and Statistics, University of Texas at San Antonio,* [c]*QuynhNhu.Nguyen@my.utsa.edu*

[3]*Department of Mathematics, Illinois State University,* [d]*mxu2@ilstu.edu*

[4]*School of Mathematics and Statistics, Jiangsu Normal University,* [e]*zhaop@jsnu.edu.cn*

Modeling cyber data breach risks poses a formidable challenge, primarily due to the intricate multivariate dependencies within a backdrop of limited data. This study proposes a novel ensemble learning approach that effectively captures both the temporal and cross-sectional dependence inherent in cyber risks. The proposed approach is significantly different from those traditional ones that directly model the multivariate dependence among risks. Instead, our approach leverages bivariate copulas to generate predictive members to capture the multivariate dependence, and the resulting predictive distribution is calibrated by minimizing the distribution score. Furthermore, the proposed model is applied in insurance pricing, and the results show that it can lead to more profitable contracts. Through extensive simulations and analysis of real-world data, our findings reveal that the proposed model has satisfactory fitting and predictive performance and outperforms those in the literature.

REFERENCES

BARAN, S. and LERCH, S. (2015). Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Q. J. R. Meteorol. Soc.* **141** 2289–2299.

BUCKMAN, J., HASHIM, M., WOUTERSEN, T. and BOCKSTEDT, J. (2018). Fool me twice: An analysis of repeat data breaches within firms. *SSRN Electron. J.* https://doi.org/10.2139/ssrn.3258599

CZADO, C. (2019). *Analyzing Dependent Data with Vine Copulas: A Practical Guide with R. Lecture Notes in Statistics* **222**. Springer, Cham. MR3931334 https://doi.org/10.1007/978-3-030-13785-4

EDWARDS, B., HOFMEYR, S. and FORREST, S. (2016). Hype and heavy tails: A closer look at data breaches. *J. Cybersecurity* **2** 3–14.

ELING, M. and JUNG, K. (2018). Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance Math. Econom.* **82** 167–180. MR3850616 https://doi.org/10.1016/j.insmatheco.2018.07.003

ELING, M. and LOPERFIDO, N. (2017). Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance Math. Econom.* **75** 126–136. MR3670067 https://doi.org/10.1016/j.insmatheco.2017.05.008

FANG, Z., XU, M., XU, S. and HU, T. (2021). A framework for predicting data breach risk: Leveraging dependence to cope with sparsity. *IEEE Trans. Inform. Forensics Secur.* **16** 2186–2201.

FREES, E. W., MEYERS, G. and CUMMINGS, A. D. (2011). Summarizing insurance scores using a Gini index. *J. Amer. Statist. Assoc.* **106** 1085–1098. MR2894766 https://doi.org/10.1198/jasa.2011.tm10506

FREES, E. W. J., MEYERS, G. and CUMMINGS, A. D. (2014). Insurance ratemaking and a Gini index. *J. Risk Insur.* **81** 335–366.

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

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

HELSKE, J. (2023). KFAS: Kalman filter and smoother for exponential family state space models. R Package Version 1.5.1.

JOE, H. (1997). *Multivariate Models and Dependence Concepts. Monographs on Statistics and Applied Probability* **73**. CRC Press, London. MR1462613 https://doi.org/10.1201/b13150

Key words and phrases. CRPS, heavy tail, Lorenz curve, vine copula.

JOE, H. (2015). *Dependence Modeling with Copulas*. *Monographs on Statistics and Applied Probability* **134**. CRC Press, Boca Raton, FL. MR3328438

KLEIBER, W., RAFTERY, A. E., BAARS, J., GNEITING, T., MASS, C. F. and GRIMIT, E. (2011). Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Weather Rev.* **139** 2630–2649.

LE, T. M. and CLARKE, B. (2022). Model averaging is asymptotically better than model selection for prediction. *J. Mach. Learn. Res.* **23** Paper No. [33], 53. MR4420758

LI, Y., NGUYEN, Q. N., XU, M., ZHAO, P. and HU, T. (2025). Supplement to "EBiCop: Ensemble Bivariate Copulas for Modeling Multivariate Cyber Data Breach Risks." https://doi.org/10.1214/24-AOAS1972SUPPA, https://doi.org/10.1214/24-AOAS1972SUPPB

MAILLART, T. and SORNETTE, D. (2010). Heavy-tailed distribution of cyber-risks. *Eur. Phys. J. B* **75** 357–364.

NAGLER, T., KRÜGER, D. and MIN, A. (2022). Stationary vine copula models for multivariate time series. *J. Econometrics* **227** 305–324. MR4384675 https://doi.org/10.1016/j.jeconom.2021.11.015

NETDILIGENCE (2022). Cyber Claim Study. Available at https://netdiligence.com/cyber-claims-study-2022-report/.

PRC (2023). Privacy Rights Clearinghouse's Chronology of Data Breaches. Available at https://www.privacyrights.org/data-breaches.

RUDOLPH, M. J. (2022). 15th Emerging Risk Survey. Available at https://www.soa.org/4a2a24/globalassets/assets/files/resources/research-report/2022/.

SCHEPSMEIER, U., STOEBER, J., BRECHMANN, E. C., GRAELER, B., NAGLER, T., ERHARDT, T., ALMEIDA, C., MIN, A., CZADO, C. et al. (2015). Package 'vinecopula'. R Package Version 2.5.0.

SECURITY, I. (2023). Cost of a Data Breach Report 2022. Available at https://www.ibm.com/downloads/cas/3R8N1DZJ.

SHI, P. and LEE, G. Y. (2022). Copula regression for compound distributions with endogenous covariates with applications in insurance deductible pricing. *J. Amer. Statist. Assoc.* **117** 1094–1109. MR4480693 https://doi.org/10.1080/01621459.2022.2040519

SHI, P. and YANG, L. (2018). Pair copula constructions for insurance experience rating. *J. Amer. Statist. Assoc.* **113** 122–133. MR3803444 https://doi.org/10.1080/01621459.2017.1330692

SLOUGHTER, M., GNEITING, T. and RAFTERY, A. E. (2010). Probabilistic wind spread forecasting using ensembles and Bayesian model averaging. *J. Amer. Statist. Assoc.* **105** 25–35. MR2757189 https://doi.org/10.1198/jasa.2009.ap08615

SUN, H., XU, M. and ZHAO, P. (2021). Modeling malicious hacking data breach risks. *N. Amer. Actuar. J.* **25** 484–502.

SUN, H., XU, M. and ZHAO, P. (2023). A multivariate frequency-severity framework for healthcare data breaches. *Ann. Appl. Stat.* **17** 240–268. MR4539030 https://doi.org/10.1214/22-aoas1625

THORARINSDOTTIR, T. L. and GNEITING, T. (2010). Probabilistic forecasts of wind speed: Ensemble model ouput statistics by using heteroscedastic censored regression. *J. Roy. Statist. Soc. Ser. A* **173** 371–388. MR2751882 https://doi.org/10.1111/j.1467-985X.2009.00616.x

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

WHEATLEY, S., MAILLART, T. and SORNETTE, D. (2016). The extreme risk of personal data breaches and the erosion of privacy. *Eur. Phys. J. B* **89** 7.

WRIGHT, J. H. (2008). Bayesian model averaging and exchange rate forecasts. *J. Econometrics* **146** 329–341. MR2465177 https://doi.org/10.1016/j.jeconom.2008.08.012

XU, M., SCHWEITZER, K. M., BATEMAN, R. M. and XU, S. (2018). Modeling and predicting cyber hacking breaches. *IEEE Trans. Inform. Forensics Secur.* **13** 2856–2871.

YEE, T. W. (2015). *Vector Generalized Linear and Additive Models*: *With an Implementation in R*. *Springer Series in Statistics*. Springer, New York. MR3408425 https://doi.org/10.1007/978-1-4939-2818-7

ZHANG, P., PITT, D. and WU, X. (2022). A new multivariate zero-inflated hurdle model with applications in automobile insurance. *Astin Bull.* **52** 393–416. MR4426579 https://doi.org/10.1017/asb.2021.39

ZHAO, Z., SHI, P. and ZHANG, Z. (2022). Modeling multivariate time series with copula-linked univariate d-vines. *J. Bus. Econom. Statist.* **40** 690–704. MR4410891 https://doi.org/10.1080/07350015.2020.1859381

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 586–613
https://doi.org/10.1214/24-AOAS1974

NESTED DIRICHLET MODELS FOR UNSUPERVISED ATTACK PATTERN DETECTION IN HONEYPOT DATA

BY FRANCESCO SANNA PASSINO[1,a], ANASTASIA MANTZIOU[2,e], DANIYAR GHANI[1,b], PHILIP THIEDE[1,c], ROSS BEVINGTON[3,f] AND NICHOLAS A. HEARD[1,d]

[1]*Department of Mathematics, Imperial College London,* [a]*f.sannapassino@imperial.ac.uk,* [b]*daniyar.ghani18@imperial.ac.uk,* [c]*thiede.philip@gmail.com,* [d]*n.heard@imperial.ac.uk*

[2]*Department of Statistics, University of Warwick,* [e]*anastasia.mantziou@warwick.ac.uk*

[3]*Microsoft Threat Intelligence Center (MSTIC),* [f]*ross.bevington@microsoft.com*

Cyber systems are under near-constant threat from intrusion attempts. Attacks types vary, but each attempt typically has a specific underlying intent, and the perpetrators are typically groups of individuals with similar objectives. Clustering attacks appearing to share a common intent is very valuable to threat-hunting experts. This article explores Dirichlet distribution topic models for clustering terminal session commands collected from honeypots, which are special network hosts designed to entice malicious attackers. The main practical implications of clustering the sessions are two-fold: finding similar groups of attacks and identifying outliers. A range of statistical models are considered, adapted to the structures of command-line syntax. In particular, concepts of primary and secondary topics, and then session-level and command-level topics, are introduced into the models to improve interpretability. The proposed methods are further extended in a Bayesian nonparametric fashion to allow unboundedness in the vocabulary size and the number of latent intents. The methods are shown to discover an unusual MIRAI variant, which attempts to take over existing cryptocurrency coin-mining infrastructure, not detected by traditional topic-modelling approaches.

REFERENCES

ADAMS, S., CARTER, B., FLEMING, C. and BELING, P. A. (2018). Selecting system specific cybersecurity attack patterns using topic modeling. In 2018 17*th IEEE International Conference on Trust*, *Security and Privacy in Computing and Communications*/12*th IEEE International Conference on Big Data Science and Engineering* 490–497.

AGGARWAL, C. C. and ZHAI, C. (2012). A survey of text classification algorithms. In *Mining Text Data* 163–222. Springer, US.

ALLISON, B., GUTHRIE, D. and GUTHRIE, L. (2006). Another look at the data sparsity problem. In *Proceedings of the* 9*th International Conference on Text*, *Speech and Dialogue*. *TSD'*06 327–334. Springer, Berlin.

ANDREW, Y., LIM, C. and BUDIARTO, E. (2022). Mapping Linux shell commands to MITRE ATT&CK using NLP-based approach. In 2022 *International Conference on Electrical Engineering and Informatics* (*ICELTICs*) 37–42.

ARCHAMBEAU, C., LAKSHMINARAYANAN, B. and BOUCHARD, G. (2015). Latent IBP compound Dirichlet allocation. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **37** 321–333. https://doi.org/10.1109/TPAMI.2014.2313122

ASCOLANI, F., LIJOI, A., REBAUDO, G. and ZANELLA, G. (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika* **110** 551–558. MR4589070 https://doi.org/10.1093/biomet/asac051

BALIKAS, G., AMINI, M.-R. and CLAUSEL, M. (2016). On a topic model for sentences. In *Proceedings of the* 39*th International ACM SIGIR Conference on Research and Development in Information Retrieval*. *SIGIR'*16 921–924. Association for Computing Machinery, New York, NY, USA.

BEVINGTON, R. (2021). Microsoft Sentinel Blog: Unusual MIRAI variant looks for mining infrastructure. https://techcommunity.microsoft.com/blog/microsoftsentinelblog/unusual-mirai-variant-looks-for-mining-infrastructure/2756669.

BIRD, S., KLEIN, E. and LOPER, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Key words and phrases. Model-based clustering, statistical cybersecurity, topic modelling.

BLEI, D. M., GRIFFITHS, T. L. and JORDAN, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* **57** Art. 7, 30. MR2606082 https://doi.org/10.1145/1667053.1667056

BLEI, D. M., JORDAN, M. I., GRIFFITHS, T. L. and TENENBAUM, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the* 16*th International Conference on Neural Information Processing Systems*. *NIPS'*03 17–24. MIT Press, Cambridge, MA, USA.

BLEI, D. M. and LAFFERTY, J. D. (2006). Dynamic topic models. In *Proceedings of the* 23*rd International Conference on Machine Learning*. *ICML'*06 113–120. Assoc. Comput. Mach., New York.

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

CAI, D., CAMPBELL, T. and BRODERICK, T. (2021). Finite mixture models do not reliably learn the number of components. In *Proceedings of the* 38*th International Conference on Machine Learning*. *Proceedings of Machine Learning Research* **139** 1158–1169. PMLR.

CAI, D., HE, X. and HAN, J. (2011). Locally consistent concept factorization for document clustering. *IEEE Trans. Knowl. Data Eng.* **23** 902–913.

CAMERLENGHI, F., DUNSON, D. B., LIJOI, A., PRÜNSTER, I. and RODRÍGUEZ, A. (2019). Latent nested nonparametric priors (with discussion). *Bayesian Anal.* **14** 1303–1356. With discussions and a rejoinder. MR4044854 https://doi.org/10.1214/19-BA1169

CHANG, J., GERRISH, S., WANG, C., BOYD-GRABER, J. and BLEI, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta, eds.) **22**. Curran Associates, Red Hook.

CRESPI, V., HARDAKER, W., ABU-EL-HAIJA, S. and GALSTYAN, A. (2021). Identifying botnet IP address clusters using natural language processing techniques on honeypot command logs. e-prints.

DESHMUKH, S., RADE, R. and KAZI, D. F. (2019). Attacker behaviour profiling using stochastic ensemble of Hidden Markov models. e-prints.

DING, R., NALLAPATI, R. and XIANG, B. (2018). Coherence-aware neural topic modeling. In *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing* 830–836. Association for Computational Linguistics, Brussels, Belgium.

DOSHI-VELEZ, F., WALLACE, B. C. and ADAMS, R. (2015). Graph-sparse LDA: A topic model with structured sparsity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. *AAAI'*15 2575–2581. AAAI Press, Menlo Park.

FAN, A., DOSHI-VELEZ, F. and MIRATRIX, L. (2019). Assessing topic model relevance: Evaluation and informative priors. *Stat. Anal. Data Min.* **12** 210–222. MR3957288 https://doi.org/10.1002/sam.11415

GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235.

HANIF, H. and MAFFEIS, S. (2022). VulBERTa: Simplified source code pre-training for vulnerability detection. In 2022 *International Joint Conference on Neural Networks* (*IJCNN*) 1–8.

HERO, A., KAR, S., MOURA, J., NEIL, J., POOR, H. V., TURCOTTE, M. and XI, B. (2023). Statistics and data science for cybersecurity. *Harv. Data Sci. Rev.* **5**.

HIGHNAM, K., ARULKUMARAN, K., HANIF, Z. and JENNINGS, N. R. (2021). *BETH Dataset*: *Real Cybersecurity Data for Anomaly Detection Research*. *ICML Workshop on Uncertainty and Robustness in Deep Learning*.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987 https://doi.org/10.1214/088342305000000016

JIANG, H., ZHOU, R., ZHANG, L., WANG, H. and ZHANG, Y. (2019). Sentence level topic models for associated topics extraction. *World Wide Web* **22** 2545–2560.

KE, Z. T. and WANG, M. (2024). Using SVD for topic modeling. *J. Amer. Statist. Assoc.* **119** 434–449. MR4713904 https://doi.org/10.1080/01621459.2022.2123813

LAFFERTY, J. and BLEI, D. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf and J. Platt, eds.) **18**. MIT Press, Cambridge.

LINGENFELTER, B., VAKILINIA, I. and SENGUPTA, S. (2020). Analyzing variation among IoT botnets using medium interaction honeypots. In 2020 10*th Annual Computing and Communication Workshop and Conference* (*CCWC*) 0761–0767.

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

MANNING, C. D., RAGHAVAN, P. and SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge.

MEDVEDOVIC, M., YEUNG, K. Y. and BUMGARNER, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20** 1222–1232. https://doi.org/10.1093/bioinformatics/bth068

MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) **26**. Curran Associates, Red Hook.

MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. MR3277163

MOKUBE, I. and ADAMS, M. (2007). Honeypots: Concepts, approaches, and challenges. In *Proceedings of the 45th Annual Southeast Regional Conference* 321–326.

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 735–749. MR2088779 https://doi.org/10.1111/j.1467-9868.2004.05564.x

NEWMAN, D., ASUNCION, A., SMYTH, P. and WELLING, M. (2009). Distributed algorithms for topic models. *J. Mach. Learn. Res.* **10** 1801–1828. MR2540777

PAISLEY, J., WANG, C., BLEI, D. M. and JORDAN, M. I. (2015). Nested hierarchical Dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 256–270. https://doi.org/10.1109/TPAMI.2014.2318728

PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

RADE, R., DESHMUKH, S., NENE, R., WADEKAR, A. S. and UNNY, A. (2018). Temporal and stochastic modelling of attacker behaviour. In *International Conference on Intelligent Information Technologies* 30–45.

ŘEHŮŘEK, R. and SOJKA, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

ROY, S., PANAOUSIS, E., NOAKES, C., LASZKA, A., PANDA, S. and LOUKAS, G. (2023). SoK: The MITRE ATT&CK Framework in Research and Practice. arXiv e-prints.

SADIQUE, F. and SENGUPTA, S. (2021). Analysis of attacker behavior in compromised hosts during command and control. In *ICC 2021—IEEE International Conference on Communications* 1–7.

SANNA PASSINO, F., MANTZIOU, A., GHANI, D., THIEDE, P., BEVINGTON, R. and HEARD, N. A (2025). Supplement to "Nested Dirichlet models for unsupervised attack pattern detection in honeypot data." https://doi.org/10.1214/24-AOAS1974SUPPA, https://doi.org/10.1214/24-AOAS1974SUPPB

SATO, I. and NAKAGAWA, H. (2010). Topic models with power-law using Pitman-yor process. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'10* 673–682. Association for Computing Machinery, New York, NY, USA.

SCHOFIELD, A., MAGNUSSON, M. and MIMNO, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* 432–436. Association for Computational Linguistics.

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

SHRIVASTAVA, R. K., BASHIR, B. and HOTA, C. (2019). Attack detection and forensics using honeypot in IoT environment. In *International Conference on Distributed Computing and Internet Technology*. 402–409.

TEH, Y., NEWMAN, D. and WELLING, M. (2006). *A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In Advances in Neural Information Processing Systems* **19**. MIT Press, Cambridge.

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

TOMASI, F., CHANDAR, P., LEVY-FIX, G., LALMAS-ROELLEKE, M. and DAI, Z. (2020). Stochastic variational inference for dynamic correlated topic models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI). Proceedings of Machine Learning Research* **124** 859–868. PMLR.

TRISTAN, J.-B., TASSAROTTI, J. and STEELE, G. (2015). Efficient training of LDA on a GPU by mean-for-mode estimation. In *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.). *Proceedings of Machine Learning Research* **37** 59–68. PMLR, Lille, France.

WALLACH, H. M. (2008). Structured topic models for language Ph.D. thesis, Univ. Cambridge.

WANG, P., WU, L., CUNNINGHAM, R. and ZOU, C. C. (2010). Honeypot detection in advanced botnet attacks. *Int. J.Inf.Comput. Secur.* **4** 30–51.

WATERS, A. S. (2014). Infinite-word topic models for digital media Ph.D. thesis, Univ. Texas at Austin.

WILLIAMSON, S., WANG, C., HELLER, K. A. and BLEI, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on International Conference on Machine Learning. ICML'10* 1151–1158. Omnipress, Madison, WI, USA.

WILSON, A. T. and CHEW, P. A. (2010). Term weighting schemes for latent Dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. HLT'10* 465–473. Association for Computational Linguistics, USA.

XIE, P. and XING, E. P. (2013). Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. UAI'13* 694–703. AUAI Press.

XU, W., LIU, X. and GONG, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. SIGIR'03* 267–273. Association for Computing Machinery, New York, NY, USA.

YAN, F., XU, N. and QI, Y. (2009). Parallel inference for latent Dirichlet allocation on graphics processing units. In *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta, eds.) **22**. Curran Associates, Red Hook.

ZHAI, K. and BOYD-GRABER, J. (2013). Online latent Dirichlet allocation with infinite vocabulary. In *Proceedings of the* 30*th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.). *Proceedings of Machine Learning Research* **28** 561–569. PMLR, Atlanta, GA, USA.

ZHANG, X. (2020). Bayesian latent feature modelling for unstructured data Ph.D. thesis, Imperial College, London.

ZHU, Y., CHEN, Z., YAN, Q., WANG, S., LI, E., PENG, L. and ZHAO, C. (2022). Mining Function Homology of Bot Loaders from Honeypot Logs. e-prints.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 614–636
https://doi.org/10.1214/24-AOAS1975

A NOVEL FRAMEWORK TO QUANTIFY UNCERTAINTY IN PEPTIDE-TANDEM MASS SPECTRUM MATCHES WITH APPLICATION TO NANOBODY PEPTIDE IDENTIFICATION

BY CHRIS MCKENNAN[1,a], ZHE SANG[2,b] AND YI SHI[2,c]

[1]*Department of Statistics, University of Pittsburgh,* [a]*chm195@pitt.edu*
[2]*Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai,* [b]*zhe.sang@gmail.com,* [c]*yi.shi@mssm.edu*

Nanobodies are small antibody fragments derived from camelids that selectively bind to antigens. These proteins have marked physicochemical properties that support advanced therapeutics, including possible treatments for SARS-CoV-2. To realize their potential, bottom-up proteomics via liquid chromatography-tandem mass spectrometry (LC-MS/MS) has been proposed to identify antigen-specific nanobodies at the proteome scale, where a critical component of this pipeline is matching nanobody peptides to their begotten tandem mass spectra. While peptide-spectrum matching is a well-studied problem, we show the sequence similarity between nanobody peptides violates key assumptions necessary to infer nanobody peptide-spectrum matches (PSMs) with the standard target–decoy paradigm and prove these violations beget inflated error rates. To address these issues, we develop a novel framework and method that treats peptide-spectrum matching as a Bayesian model selection problem with an incomplete model space, which are, to our knowledge, the first to account for all sources of PSM error without relying on the aforementioned assumptions. Our work also demonstrates how to leverage novel retention time and spectrum prediction tools to develop accurate and discriminating data-generating models and, to our knowledge, provides the first rigorous description of MS/MS spectrum noise. We illustrate our method's superior performance on simulated and real nanobody data.

REFERENCES

BAGAL, D., KAST, E. and CAO, P. (2017). Rapid distinction of leucine and isoleucine in monoclonal antibodies using nanoflow LCMS^n. *Anal. Chem.* **89** 720–727.

BANNAS, P., HAMBACH, J. and KOCH-NOLTE, F. (2017). Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics. *Front. Immunol.* **8**. https://doi.org/10.3389/fimmu.2017.01603

COX, J., NEUHAUSER, N., MICHALSKI, A., SCHELTEMA, R. A., OLSEN, J. V. and MANN, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10** 1794–1805. https://doi.org/10.1021/pr101065j

CRAIG, R. and BEAVIS, R. C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17** 2310–2316. https://doi.org/10.1002/rcm.1198

DORFER, V., PICHLER, P., STRANZL, T., STADLMANN, J., TAUS, T., WINKLER, S. and MECHTLER, K. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **13** 3679–3684. https://doi.org/10.1021/pr500202e

ELIAS, J. E. and GYGI, S. P. (2007). Target–decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4** 207–214. https://doi.org/10.1038/nmeth1019

ENG, J. K., MCCORMACK, A. L. and YATES, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Amer. Soc. Mass Spectrom.* **5** 976–989.

FENG, X.-D., LI, L.-W., ZHANG, J.-H., ZHU, Y.-P., CHANG, C., SHU, K.-X. and MA, J. (2017). Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC Genomics* **18** 143–143.

Key words and phrases. Bayesian model selection, local false discovery rate, mass spectrometry, nanobody, peptide-spectrum match, proteomics.

FRIDY, P. C., LI, Y., KEEGAN, S., THOMPSON, M. K., NUDELMAN, I., SCHEID, J. F., OEFFINGER, M., NUSSENZWEIG, M. C., FENYÖ, D. et al. (2014). A robust pipeline for rapid production of versatile nanobody repertoires. *Nat. Methods* **11** 1253–1260. https://doi.org/10.1038/nmeth.3170

GESSULAT, S., SCHMIDT, T., ZOLG, D. P., SAMARAS, P., SCHNATBAUM, K., ZERWECK, J., KNAUTE, T., RECHENBERGER, J., DELANGHE, B. et al. (2019). Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16** 509–518. https://doi.org/10.1038/s41592-019-0426-7

HAMERS-CASTERMAN, C., ATARHOUCH, T., MUYLDERMANS, S., ROBINSON, G., HAMMERS, C., SONGA, E. B., BENDAHMAN, N. and HAMMERS, R. (1993). Naturally occurring antibodies devoid of light chains. *Nature* **363** 446–448. https://doi.org/10.1038/363446a0

HOWBERT, J. J. and NOBLE, W. S. (2014). Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **13** 2467–2479.

JOVČEVSKA, I. and MUYLDERMANS, S. (2020). The therapeutic potential of nanobodies. *BioDrugs* **34** 11–26.

KÄLL, L., STOREY, J. D. and NOBLE, W. S. (2008). Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **24** i42–i48. https://doi.org/10.1093/bioinformatics/btn294

KÄLL, L., STOREY, J. D. and NOBLE, W. S. (2009). QVALITY: Non-parametric estimation of q-values and posterior error probabilities. *Bioinformatics* **25** 964–966. https://doi.org/10.1093/bioinformatics/btp021

KÄLL, L. and VITEK, O. (2011). Computational mass spectrometry-based proteomics. *PLoS Comput. Biol.* **7** e1002277. https://doi.org/10.1371/journal.pcbi.1002277

KEICH, U., KERTESZ-FARKAS, A. and NOBLE, W. S. (2015). Improved false discovery rate estimation procedure for shotgun proteomics. *J. Proteome Res.* **14** 3148–3161.

KEICH, U. and NOBLE, W. S. (2018). Controlling the FDR in imperfect matches to an incomplete database. *J. Amer. Statist. Assoc.* **113** 973–982. MR3862332 https://doi.org/10.1080/01621459.2017.1375931

KIM, S. and PEVZNER, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5** 5277. https://doi.org/10.1038/ncomms6277

LEVENSHTEIN, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.* **10** 707–710.

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

MCKENNAN, C., SANG, Z. and SHI, Y. (2025). Supplement to "A novel framework to quantify uncertainty in peptide-tandem mass spectrum matches with application to nanobody peptide identification." https://doi.org/10.1214/24-AOAS1975SUPPA, https://doi.org/10.1214/24-AOAS1975SUPPB

MITCHELL, L. S. and COLWELL, L. J. (2018). Comparative analysis of nanobody sequence and structure data. *Proteins, Struct. Funct. Bioinform.* **86** 697–706. https://doi.org/10.1002/prot.25497

ORSBURN, B. C. (2021). Proteome discoverer—a community enhanced data processing suite for protein informatics. *Proteomes* **9** 15. https://doi.org/10.3390/proteomes9010015

PERKINS, D. N., PAPPIN, D. J., CREASY, D. M. and COTTRELL, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20** 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

SCHOENHOLZ, S. S., HACKETT, S., DEMING, L., MELAMUD, E., JAITLY, N., MCALLISTER, F., O'BRIEN, J., DAHL, G., BENNETT, B. et al. (2018). Peptide-Spectra Matching from Weak Supervision.

SHEN, Z., SANG, Z. and SHI, Y. (2022). Nanobodies as a powerful platform for biomedicine. *Trends Mol. Med.* **28** 1006–1007. https://doi.org/10.1016/j.molmed.2022.08.007

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

THE, M., MACCOSS, M. J., NOBLE, W. S. and KÄLL, L. (2016). Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *J. Amer. Soc. Mass Spectrom.* **27** 1719–1727. https://doi.org/10.1007/s13361-016-1460-7

TYANOVA, S., TEMU, T. and COX, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11** 2301–2319. https://doi.org/10.1038/nprot.2016.136

WAN, Y., YANG, A. and CHEN, T. (2006). PepHMM: A hidden Markov model based scoring function for mass spectrometry database search. *Anal. Chem.* **78** 432–437.

WEBB, K., BRISTOW, A., SARGENT, M. and STEIN, B. (2004). *Best Practice Guide: Methodology for Accurate Mass Measurement of Small Molecules*. LGC Ltd., London.

XIANG, Y., NAMBULLI, S., XIAO, Z., LIU, H., SANG, Z., DUPREX, W. P., SCHNEIDMAN-DUHOVNY, D., ZHANG, C. and SHI, Y. (2020). Versatile and multivalent nanobodies efficiently neutralize SARS-CoV-2. *Science* **370** 1479–1484. https://doi.org/10.1126/science.abe4747

XIANG, Y., SANG, Z., BITTON, L., XU, J., LIU, Y., SCHNEIDMAN-DUHOVNY, D. and SHI, Y. (2021). Integrative proteomics identifies thousands of distinct, multi-epitope, and high-affinity nanobodies. *Cell Syst.* **12** 220–234.e9. https://doi.org/10.1016/j.cels.2021.01.003

Xu, J. L. and Davis, M. M. (2000). Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13** 37–45. https://doi.org/10.1016/s1074-7613(00)00006-6

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 637–654
https://doi.org/10.1214/24-AOAS1976

HETEROGENEOUS NETWORK ANALYSIS OF DISEASE CLINICAL TREATMENT MEASURES VIA MINING ELECTRONIC MEDICAL RECORD DATA

BY JIPING WANG[1,a], RONG LI[1,b], WEI-SHAN CHANG[2,d], KAI-YUAN HSIAO[2,e], BEN-CHANG SHIA[2,f] AND SHUANGGE MA[1,c]

[1]*Department of Biostatistics, Yale School of Public Health,* [a]*jiping.wang@yale.edu,* [b]*rong.li.rl946@yale.edu,* [c]*shuangge.ma@yale.edu*

[2]*Graduate Institute of Business Administration, College of Management; Artificial Intelligence Development Center, Fu Jen Catholic University,* [d]*sanchangai@gmail.com,* [e]*st880005@gmail.com,* [f]*025674@mail.fju.edu.tw*

The analysis of clinical treatment measures has been extensively conducted and can facilitate more effective resource management and assist in better understanding diseases. Most of the existing analyses have been focused on a single disease or many diseases combined. Partly motivated by the successes of gene-centric and phenotypic human disease network (HDN) research, there has been growing interest in network analysis of clinical treatment measures. However, existing studies have been limited by a lack of attention to heterogeneity and relevant covariates, ineffectiveness of methods, and low data quality. In this study our goal is to mine the Taiwan National Health Insurance Research Database (NHIRD), a large population-level electronic medical record (EMR) database, and construct HDNs for the number of outpatient visits and medical cost. Significantly advancing from existing literature, the proposed analysis accommodates heterogeneity and the effects of covariates. Additionally, the proposed method effectively accommodates zero inflation, Poisson distribution, high dimensionality, and network sparsity. Computational and theoretical properties are carefully examined. Simulation demonstrates the competitive performance of the proposed approach. In the analysis of NHIRD data, two and four subject groups are identified for outpatient visit and medical cost, respectively. The interconnections, hubs, and network modules are found to have sound implications.

REFERENCES

BRUGNARA, G., NEUBERGER, U., MAHMUTOGLU, M. A., FOLTYN, M., HERWEH, C., NAGEL, S., SCHÖNENBERGER, S., HEILAND, S., ULFERT, C. et al. (2020). Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke* **51** 3541–3551.

CHANG, X., HUANG, D. and WANG, H. (2019). A popularity-scaled latent space model for large-scale directed social network. *Statist. Sinica* **29** 1277–1299. MR3932518

CHIQUET, J., MARIADASSOU, M. and ROBIN, S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Front. Ecol. Evol.* **9** 588292.

CHIQUET, J., ROBIN, S. and MARIADASSOU, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning* 1162–1171. PMLR.

CLANCY, C. M. and EISENBERG, J. M. (1998). Outcomes research: Measuring the end results of health care. *Science* **282** 245–246. https://doi.org/10.1126/science.282.5387.245

CURRY, L. A., NEMBHARD, I. M. and BRADLEY, E. H. (2009). Qualitative and mixed methods provide unique contributions to outcomes research. *Circulation* **119** 1442–1452.

FAREWELL, V. T., LONG, D. L., TOM, B. D. M., YIU, S. and SU, L. (2017). Two-part and related regression models for longitudinal data. *Annu. Rev. Stat. Appl.* **4** 283–315. https://doi.org/10.1146/annurev-statistics-060116-054131

GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in Gaussian mixture models. *Electron. J. Stat.* **10** 1133–1154. MR3499523 https://doi.org/10.1214/16-EJS1135

Key words and phrases. Clinical treatment measure, human disease network, heterogeneity, NHIRD.

GOH, K.-I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. and BARABÁSI, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690.

HIDALGO, C. A., BLUMM, N., BARABÁSI, A.-L. and CHRISTAKIS, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* **5** e1000353.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HORVATH, S. (2011). *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer, Berlin.

HORWITZ, L. I., WANG, Y., DESAI, M. M., CURRY, L. A., BRADLEY, E. H., DRYE, E. E. and KRUMHOLZ, H. M. (2012). Correlations among risk-standardized mortality rates and among risk-standardized readmission rates within hospitals. *J. Hosp. Med.* **7** 690–696. https://doi.org/10.1002/jhm.1965

LAI, K., CUI, J., NI, S., ZHANG, Y., HE, J. and YAO, K. (2013). The effects of postmenopausal hormone use on cataract: A meta-analysis. *PLoS ONE* **8** e78647.

LEE, J.-S., CHEN, W.-M., HUANG, L.-H., CHUNG, C.-C., YU, K.-H., KUO, C.-F. and SEE, L.-C. (2020). Epidemiology of outpatient and inpatient eye injury in Taiwan: 2000, 2005, 2010, and 2015. *PLoS ONE* **15** e0235208. https://doi.org/10.1371/journal.pone.0235208

LI, B., CHUN, H. and ZHAO, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Amer. Statist. Assoc.* **107** 152–167. MR2949348 https://doi.org/10.1080/01621459.2011.644498

MA, C., LI, Y., SHIA, B. and MA, S. (2020). Human disease cost network analysis. *Stat. Med.* **39** 1237–1249. MR4098487 https://doi.org/10.1002/sim.8472

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

MCPHAIL, S. M. (2016). Multimorbidity in chronic disease: Impact on health care resources and costs. *Risk. Manag. Healthc. Policy* **9** 143–156. https://doi.org/10.2147/RMHP.S97248

MEI, H., JIA, R., QIAO, G., LIN, Z. and MA, S. (2021). Human disease clinical treatment network for the elderly: The analysis of medicare inpatient length of stay data. *Stat. Med.* **40** 2083–2099. MR4229840 https://doi.org/10.1002/sim.8893

MEI, H., JIA, R., QIAO, G., LIN, Z. and MA, S. (2023). Human disease clinical treatment network for the elderly: Analysis of the medicare inpatient length of stay and readmission data. *Biometrics* **79** 404–416. MR4572531 https://doi.org/10.1111/biom.13549

NORAN, N., SALLEH, N. and ZAHARI, M. (2007). Relationship between reproductive exposures and age-related cataract in women. *Asia-Pac. J. Public Health* **19** 23–28.

REN, M., ZHANG, S., ZHANG, Q. and MA, S. (2022). Gaussian graphical model-based heterogeneity analysis via penalized fusion. *Biometrics* **78** 524–535. MR4450573 https://doi.org/10.1111/biom.13426

SHRANK, W. H., ROGSTAD, T. L. and PAREKH, N. (2019). Waste in the US health care system: Estimated costs and potential for savings. *JAMA* **322** 1501–1509. https://doi.org/10.1001/jama.2019.13978

SUN, K., GONÇALVES, J. P., LARMINIE, C. and PRŽULJ, N. (2014). Predicting disease associations via biological network analysis. *BMC Bioinform.* **15** 1–13.

TSUGAWA, Y., JENA, A. B., FIGUEROA, J. F., ORAV, E. J., BLUMENTHAL, D. M. and JHA, A. K. (2017). Comparison of hospital mortality and readmission rates for medicare patients treated by male vs female physicians. *JAMA Intern. Med.* **177** 206–213. https://doi.org/10.1001/jamainternmed.2016.7875

TUMEH, P. C., HELLMANN, M. D., HAMID, O., TSAI, K. K., LOO, K. L., GUBENS, M. A., ROSENBLUM, M., HARVIEW, C. L., TAUBE, J. M. et al. (2017). Liver metastasis and treatment outcome with anti-PD-1 monoclonal antibody in patients with melanoma and NSCLC. *Cancer Immunol. Res.* **5** 417–424.

WANG, J., LI, R., CHANG, W.-S., HSIAO, K.-Y., SHIA, B.-C. and MA, S. (2025). Supplement to “Heterogeneous Network Analysis of Disease Clinical Treatment Measures via Mining Electronic Medical Record Data.” https://doi.org/10.1214/24-AOAS1976SUPPA, https://doi.org/10.1214/24-AOAS1976SUPPB, https://doi.org/10.1214/24-AOAS1976SUPPC, https://doi.org/10.1214/24-AOAS1976SUPPD, https://doi.org/10.1214/24-AOAS1976SUPPE

ZETTERBERG, M. (2016). Age-related eye disease and gender. *Maturitas* **83** 19–26. https://doi.org/10.1016/j.maturitas.2015.10.005

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 655–679
https://doi.org/10.1214/24-AOAS1978

VARIABLE SCREENING AND SPATIAL SMOOTHING IN FRÉCHET REGRESSION WITH APPLICATION TO DIFFUSION TENSOR IMAGING

BY LEI YAN[1,a], XIN ZHANG[1,b], ZHOU LAN[2,c], DIPANKAR BANDYOPADHYAY[3,d], YICHAO WU[4,e] AND FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

[1]*Department of Statistics, Florida State University,* [a]*lyan@fsu.edu,* [b]*henry@stat.fsu.edu*
[2]*Department of Radiology, Brigham and Women's Hospital and Harvard Medical School,* [c]*zlan@bwh.harvard.edu*
[3]*Department of Biostatistics, Virginia Commonwealth University,* [d]*dbandyop@vcu.edu*
[4]*Department of Mathematics, Statistics, and Computer Science, University of Illinois,* [e]*yichaowu@uic.edu*

Modern applications in medical imaging often include high-dimensional predictors and spatially dependent responses in the non-Euclidean space. For example, in imaging-genetics studies, our objective is to study the relationship between single-nucleotide polymorphisms (SNPs), a high-dimensional predictor vector, and diffusion tensor imaging (DTI) responses, which are thousands to millions of voxelwise 3×3 symmetric positive definite (SPD) matrices. In this paper we develop a fast and pragmatic method of regressing spatially associated random responses on a high-dimensional predictor set. Specifically, we focus on two related problems: fast variable screening of high-dimensional predictors and smoothing techniques for non-Euclidean spatially associated responses. Under a Fréchet regression framework (which handles regression of SPD matrix-variate responses on covariates in Euclidean space), we propose a two-stage approach, where a screening method (using distance covariances in metric spaces) is employed to mitigate high-dimensionality (Stage 1), followed by deriving a closed-form solution that powers elegant smoothing of the spatially associated SPD responses (Stage 2). We investigate the finite-sample properties of our method using synthetic data generated under various settings and present illustration via analysis of an imaging-genetics (DTI responses with genetic and demographic predictors) dataset, derived from the Alzheimer's Disease Neuroimaging Initiative 2. Code for implementing our proposed method is available in the `GitHub` link: https://github.com/leiyan-ly/Frechet-regression.

REFERENCES

ALEXANDER, A. L., LEE, J. E., LAZAR, M. and FIELD, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics* **4** 316–329.

ASSAREH, A. A., PIGUET, O., LYE, T. C., MATHER, K. A., BROE, G. A., SCHOFIELD, P. R., SACHDEV, P. S. and KWOK, J. B. J. (2014). Association of SORL1 gene variants with hippocampal and cerebral atrophy and Alzheimer's disease. *Curr. Alzheimer Res.* **11** 558–563. https://doi.org/10.2174/1567205011666140618101408

BAI, F., YUAN, Y., SHI, Y. and ZHANG, Z. (2016). Multiple genetic imaging study of the association between cholesterol metabolism and brain functional alterations in individuals with risk factors for Alzheimer's disease. *Oncotarget* **7** 15315.

CARMICHAEL, O., CHEN, J., PAUL, D. and PENG, J. (2013). Diffusion tensor smoothing through weighted Karcher means. *Electron. J. Stat.* **7** 1913–1956. MR3084676 https://doi.org/10.1214/13-EJS825

CHAUDHURI, A. and HU, W. (2019). A fast algorithm for computing distance correlation. *Comput. Statist. Data Anal.* **135** 15–24. MR3913146 https://doi.org/10.1016/j.csda.2019.01.016

CHOWDHURY, M. Z. I. and TURIN, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Fam. Med. Commun. Health* **8** e000262. https://doi.org/10.1136/fmch-2019-000262

CUI, H., LI, R. and ZHONG, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Amer. Statist. Assoc.* **110** 630–641. MR3367253 https://doi.org/10.1080/01621459.2014.920256

Key words and phrases. Diffusion tensor imaging, Fréchet regression, distance covariance, sure screening property, smoothing.

DEMIRHAN, A., NIR, T. M., ZAVALIANGOS-PETROPULU, A., JACK, C. R., WEINER, M. W., BERNSTEIN, M. A., THOMPSON, P. M. and JAHANSHAD, N. (2015). Feature selection improves the accuracy of classifying Alzheimer disease using diffusion tensor images. In 2015 *IEEE* 12*th International Symposium on Biomedical Imaging* (*ISBI*) 126–130. IEEE.

DEZA, M. M. and DEZA, E. (2014). *Encyclopedia of Distances*, 3rd ed. Springer, Heidelberg. MR3243690 https://doi.org/10.1007/978-3-662-44342-2

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

DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821. MR4031200 https://doi.org/10.1093/biomet/asz052

ENNIS, D. B. and KINDLMANN, G. (2006). Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magn. Reson. Med.* **55** 136–146.

ESRAEL, S. M. A. M., HAMED, A. M. M., KHEDR, E. M. and SOLIMAN, R. K. (2021). Application of diffusion tensor imaging in Alzheimer's disease: Quantification of white matter microstructural changes. *Egypt. J. Radiol. Nucl. Med.* **52** 1–8.

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

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

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

GRAÑA, M., TERMENON, M., SAVIO, A., GONZALEZ-PINTO, A., ECHEVESTE, J., PÉREZ, J. M. and BESGA, A. (2011). Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation. *Neurosci. Lett.* **502** 225–229. https://doi.org/10.1016/j.neulet.2011.07.049

GREENLAW, K., SZEFER, E., GRAHAM, J., LESPERANCE, M., NATHOO, F. S. and INITIATIVE, A. D. N. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* **33** 2513–2522.

GUERREIRO, R. and BRAS, J. (2015). The age factor in Alzheimer's disease. *Gen. Med.* **7** 1–3.

HAN, F. and PAN, W. (2012). A composite likelihood approach to latent multivariate Gaussian modeling of SNP data with application to genetic association testing. *Biometrics* **68** 307–315. MR2909887 https://doi.org/10.1111/j.1541-0420.2011.01649.x

HEINZE, G., WALLISCH, C. and DUNKLER, D. (2018). Variable selection—A review and recommendations for the practicing statistician. *Biom. J.* **60** 431–449. MR3802914 https://doi.org/10.1002/bimj.201700067

HU, Y., CHENG, L., ZHANG, Y., BAI, W., ZHOU, W., WANG, T., HAN, Z., ZONG, J., JIN, S. et al. (2017). Rs4878104 contributes to Alzheimer's disease risk and regulates DAPK1 gene expression. *Neurol. Sci.* **38** 1255–1262.

JAKOBSEN, M. E. (2017). Distance covariance in metric spaces: Non-parametric independence testing in metric spaces. Master's Thesis, Univ. Copenhagen.

JEON, J. M., PARK, B. U. and VAN KEILEGOM, I. (2021). Additive regression for non-Euclidean responses and predictors. *Ann. Statist.* **49** 2611–2641. MR4338377 https://doi.org/10.1214/21-aos2048

JOHNSON, T. D., LIU, Z., BARTSCH, A. J. and NICHOLS, T. E. (2013). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Stat. Methods Med. Res.* **22** 364–381. MR3190664 https://doi.org/10.1177/0962280212448970

KALOUSIS, A., PRADOS, J. and HILARIO, M. (2005). Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining* (*ICDM'*05) 8 pp. IEEE.

KIRILINA, E., HELBLING, S., MORAWSKI, M., PINE, K., REIMANN, K., JANKUHN, S., DINSE, J., DEISTUNG, A., REICHENBACH, J. R. et al. (2020). Superficial white matter imaging: Contrast mechanisms and whole-brain in vivo mapping. *Sci. Adv.* **6** eaaz9281.

KITTS, A. and SHERRY, S. (2014). The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. In *The NCBI Handbook* [*Internet*] (J. McEntyre and J. Ostell, eds.) 266–290 5. National Center for Biotechnology Information (US), Bethesda, MD.

KRONLAGE, M., SCHWEHR, V., SCHWARZ, D., GODEL, T., UHLMANN, L., HEILAND, S., BENDSZUS, M. and BÄUMER, P. (2018). Peripheral nerve diffusion tensor imaging (DTI): Normal values and demographic determinants in a cohort of 60 healthy individuals. *Eur. Radiol.* **28** 1801–1808.

LAN, Z., REICH, B. J. and BANDYOPADHYAY, D. (2021). A spatial Bayesian semiparametric mixture model for positive definite matrices with applications in diffusion tensor imaging. *Canad. J. Statist.* **49** 129–149. MR4238358 https://doi.org/10.1002/cjs.11601

LAN, Z., REICH, B. J., GUINNESS, J., BANDYOPADHYAY, D., MA, L. and MOELLER, F. G. (2022). Geostatistical modeling of positive-definite matrices: An application to diffusion tensor imaging. *Biometrics* **78** 548–559. MR4450575 https://doi.org/10.1111/biom.13445

LEE, H. N. and SCHWARTZMAN, A. (2017). Inference for eigenvalues and eigenvectors in exponential families of random symmetric matrices. *J. Multivariate Anal.* **162** 152–171. MR3719341 https://doi.org/10.1016/j.jmva.2017.08.006

LEE, J. D., SUN, Y. and TAYLOR, J. E. (2015). On model selection consistency of regularized M-estimators. *Electron. J. Stat.* **9** 608–642. MR3331852 https://doi.org/10.1214/15-EJS1013

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900 https://doi.org/10.1080/01621459.2012.695654

LIN, Z., MÜLLER, H.-G. and PARK, B. U. (2023). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika* **110** 361–379. MR4588358 https://doi.org/10.1093/biomet/asac055

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562 https://doi.org/10.1080/01621459.2013.850086

LO, C.-Y., WANG, P.-N., CHOU, K.-H., WANG, J., HE, Y. and LIN, C.-P. (2010). Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer's disease. *J. Neurosci.* **30** 16876–16885.

LUSTGARTEN, J. L., GOPALAKRISHNAN, V. and VISWESWARAN, S. (2009). Measuring stability of feature selection in biomedical datasets. In *AMIA Annual Symposium Proceedings* **2009** 406. American Medical Informatics Association.

LYONS, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305. MR3127883 https://doi.org/10.1214/12-AOP803

MCCARTHY, J. J., SAITH, S., LINNERTZ, C., BURKE, J. R., HULETTE, C. M., WELSH-BOHMER, K. A. and CHIBA-FALEK, O. (2012). The Alzheimer's associated 5' region of the SORL1 gene cis regulates SORL1 transcripts expression. *Neurobiol. Aging* **33** 1485.e1–1485.e14858. https://doi.org/10.1016/j.neurobiolaging.2010.10.004

MIN, K. and MAI, Q. (2022). A general framework for tensor screening through smoothing. *Electron. J. Stat.* **16** 451–497. MR4361747 https://doi.org/10.1214/21-ejs1954

MORI, S. (2007). *Introduction to Diffusion Tensor Imaging*. Elsevier, Amsterdam.

NOGUEIRA, S., SECHIDIS, K. and BROWN, G. (2017). On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **18** Paper No. 174, 54 pp. MR3827062

O'DONNELL, L. J. and WESTIN, C.-F. (2011). An introduction to diffusion tensor image analysis. *Neurosurg. Clin.* **22** 185–196.

ORSI, F. (2023). Exploring the role of compactness in path-dependent land-taking processes in Italy. *GeoJournal* **88** 69–87.

PAN, W., WANG, X., XIAO, W. and ZHU, H. (2019). A generic sure independence screening procedure. *J. Amer. Statist. Assoc.* **114** 928–937. MR3963192 https://doi.org/10.1080/01621459.2018.1462709

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

REICH, B. J., GUINNESS, J., VANDEKAR, S. N., SHINOHARA, R. T. and STAICU, A.-M. (2018). Fully Bayesian spectral methods for imaging data. *Biometrics* **74** 645–652. MR3825351 https://doi.org/10.1111/biom.12782

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

SCHWARTZMAN, A. (2016). Lognormal distributions and geometric averages of symmetric positive definite matrices. *Int. Stat. Rev.* **84** 456–486. MR3580425 https://doi.org/10.1111/insr.12113

SOARES, J. M., MARQUES, P., ALVES, V. and SOUSA, N. (2013). A hitchhiker's guide to diffusion tensor imaging. *Front. Neurosci.* **7** 31.

SPENCE, J. S., CARMACK, P. S., GUNST, R. F., SCHUCANY, W. R., WOODWARD, W. A. and HALEY, R. W. (2007). Accounting for spatial dependence in the analysis of SPECT brain imaging data. *J. Amer. Statist. Assoc.* **102** 464–473. MR2370847 https://doi.org/10.1198/016214506000001284

STEBBINS, G. and MURPHY, C. (2009). Diffusion tensor imaging in Alzheimer's disease and mild cognitive impairment. *Behav. Neurol.* **21** 39–49.

SZEFER, E., LU, D., NATHOO, F., BEG, M. F. and GRAHAM, J. (2017). Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: Discovery, refinement and validation. *Stat. Appl. Genet. Mol. Biol.* **16** 367–386. MR3725750 https://doi.org/10.1515/sagmb-2016-0077

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

TUCKER, D. C., WU, Y. and MÜLLER, H.-G. (2023). Variable selection for global Fréchet regression. *J. Amer. Statist. Assoc.* **118** 1023–1037. MR4595474 https://doi.org/10.1080/01621459.2021.1969240

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of FMRI data. *IEEE Trans. Med. Imag.* **23** 213–231. https://doi.org/10.1109/TMI.2003.823065

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

XUE, W., BOWMAN, F. D. and KANG, J. (2018). A Bayesian spatial model to predict disease status using imaging data from various modalities. *Front*. *Neurosci*. **12** 184.

YAN, L., ZHANG, X., LAN, Z., BANDYOPADHYAY, D., WU, Y. and FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2025). Supplement to "Variable screening and spatial smoothing in Fréchet regression with application to diffusion tensor imaging." https://doi.org/10.1214/24-AOAS1978SUPP

ZHANG, Q., XUE, L. and LI, B. (2024). Dimension reduction for Fréchet regression. *J*. *Amer*. *Statist*. *Assoc*. **119** 2733–2747. MR4833911 https://doi.org/10.1080/01621459.2023.2277406

ZHU, Y., BLUMENTHAL, W. and LOWE, T. (1997). Determination of non-symmetric 3-D fiber-orientation distribution and average fiber length in short-fiber composites. *J*. *Compos*. *Mater*. **31** 1287–1301.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 680–700
https://doi.org/10.1214/24-AOAS1979
This research was funded, in whole or in part, by [SNSF, 200020_207367]. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions.

ADDITIVE DENSITY-ON-SCALAR REGRESSION IN BAYES HILBERT SPACES WITH AN APPLICATION TO GENDER ECONOMICS

BY EVA-MARIA MAIER[1,a], ALMOND STÖCKER[1,2,c], BERND FITZENBERGER[3,d] AND SONJA GREVEN[1,b]

[1]*School of Business and Economics, Humboldt-Universität zu Berlin,* [a]*eva-maria.maier@hu-berlin.de,* [b]*sonja.greven@hu-berlin.de*

[2]*Institute of Mathematics, EPFL,* [c]*almond.stoecker@epfl.ch*

[3]*IAB (Institute for Employment Research) and School of Business, Economics and Society, FAU Erlangen-Nürnberg ,* [d]*bernd.fitzenberger@fau.de*

Motivated by research on gender identity norms and the distribution of the woman's share in a couple's total labor income, we consider additive regression models for densities as responses with scalar covariates. To preserve nonnegativity and integration to one under vector space operations, we formulate the model for densities in a Bayes Hilbert space, which allows to not only consider continuous densities but also, for example, discrete or mixed densities. Mixed ones occur in our application, as the woman's income share is a continuous variable having discrete point masses at zero and one for single-earner couples. Estimation is based on a gradient boosting algorithm, allowing for potentially numerous flexible (linear, nonlinear, categorical, interaction etc.) covariate effects and model selection. We show useful properties of Bayes Hilbert spaces related to subcompositional coherence, also yielding new (odds-ratio) interpretations of effect functions and simplified estimation for mixed densities via an orthogonal decomposition. Applying our approach to data from the German Socio-Economic Panel Study (SOEP) shows a more symmetric distribution in East German than in West German couples after reunification and a smaller child penalty comparing couples with and without minor children. These West–East differences become smaller but are persistent over time.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647

BEBLO, M. and GÖRGES, L. (2018). On the nature of nurture. The malleability of gender differences in work preferences. *J. Econ. Behav. Organ.* **151** 19–41.

BERTRAND, M., KAMENICA, E. and PAN, J. (2015). Gender identity and relative income within households. *Q. J. Econ.* **130** 571–614.

BLACK, S. E. and SPITZ-OENER, A. (2010). Explaining women's success: Technological change and the skill content of women's work. *Rev. Econ. Stat.* **92** 187–194.

BROCKHAUS, S. and RÜGAMER, D. (2018). FDboost: Boosting Functional Regression Models R package version 0.3-2.

BROCKHAUS, S., RÜGAMER, D. and GREVEN, S. (2020). Boosting functional regression models with FDboost. *J. Stat. Softw.* **94** 1–50.

BROCKHAUS, S., SCHEIPL, F., HOTHORN, T. and GREVEN, S. (2015). The functional linear array model. *Stat. Model.* **15** 279–300. MR3349797 https://doi.org/10.1177/1471082X14566913

BÜHLMANN, P. and YU, B. (2003). Boosting with the L_2 loss: Regression and classification. *J. Amer. Statist. Assoc.* **98** 324–339. MR1995709 https://doi.org/10.1198/016214503000125

CHEN, S. X. (1999). Beta kernel estimators for density functions. *Comput. Statist. Data Anal.* **31** 131–145. MR1718494 https://doi.org/10.1016/S0167-9473(99)00010-9

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and MELLY, B. (2013). Inference on counterfactual distributions. *Econometrica* **81** 2205–2268. MR3138546 https://doi.org/10.3982/ECTA10582

Key words and phrases. Density regression, functional additive model, gradient boosting, mixed densities.

CORTES, P. and PAN, J. (2018). Occupation and gender. In *The Oxford Handbook of Women and the Economy* 425–452.

DABO-NIANG, S. and RHOMARI, N. (2009). Kernel regression estimation in a Banach space. *J. Statist. Plann. Inference* **139** 1421–1434. MR2485136 https://doi.org/10.1016/j.jspi.2008.06.015

EGOZCUE, J. J. (2009). Reply to "On the Harker variation diagrams;..." by JA Cortés. *Math. Geosci.* **41** 829–834.

EGOZCUE, J. J., DÍAZ-BARRERO, J. L. and PAWLOWSKY-GLAHN, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Math. Sin.* (*Engl. Ser.*) **22** 1175–1182. MR2245249 https://doi.org/10.1007/s10114-005-0678-2

EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2011). Basic concepts and procedures. In *Compositional Data Analysis*: *Theory and Applications* (V. Pawlowsky-Glahn and A. Buccianti, eds.) 12–28. Wiley, New York.

FITZENBERGER, B., SOMMERFELD, K. and STEFFES, S. (2013). Causal effects on employment after first birth—a dynamic treatment approach. *Labour Econ.* **25** 49–62.

GHODRATI, L. and PANARETOS, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika* **109** 957–974. MR4519110 https://doi.org/10.1093/biomet/asac005

GOEBEL, J., GRABKA, M. M., LIEBIG, S., KROH, M., RICHTER, D., SCHRÖDER, C. and SCHUPP, J. (2019). The German socio-economic panel (SOEP). *Jahrb. Natl.ökon. Stat.* **239** 345–360.

GU, C. (1995). Smoothing spline density estimation: Conditional distribution. *Statist. Sinica* **5** 709–726. MR1347615

HALL, P., WOLFF, R. C. L. and YAO, Q. (1999). Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* **94** 154–163. MR1689221 https://doi.org/10.2307/2669691

HAN, K., MÜLLER, H.-G. and PARK, B. U. (2020). Additive functional regression for densities as responses. *J. Amer. Statist. Assoc.* **115** 997–1010. MR4107695 https://doi.org/10.1080/01621459.2019.1604365

HOFNER, B., HOTHORN, T., KNEIB, T. and SCHMID, M. (2011). A framework for unbiased model selection based on boosting. *J. Comput. Graph. Statist.* **20** 956–971. MR2878957 https://doi.org/10.1198/jcgs.2011.09220

HOTHORN, T., KNEIB, T. and BÜHLMANN, P. (2014). Conditional transformation models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 3–27. MR3153931 https://doi.org/10.1111/rssb.12017

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

JEON, J. M. and PARK, B. U. (2020). Additive regression with Hilbertian responses. *Ann. Statist.* **48** 2671–2697. MR4152117 https://doi.org/10.1214/19-AOS1902

KLEVEN, H., LANDAIS, C., POSCH, J., STEINHAUER, A. and ZWEIMÜLLER, J. (2019). Child penalties across countries: Evidence and explanations. *AEA Pap. Proc.* **109** 122–26.

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

KUEHNLE, D., OBERFICHTNER, M. and OSTERMANN, K. (2021). Revisiting gender identity and relative income within households: A cautionary tale on the potential pitfalls of density estimators. *J. Appl. Econometrics* **36** 1065–1073. MR4362607 https://doi.org/10.1002/jae.2853

LI, R., REICH, B. J. and BONDELL, H. D. (2021). Deep distribution regression. *Comput. Statist. Data Anal.* **159** Paper No. 107203, 11. MR4233348 https://doi.org/10.1016/j.csda.2021.107203

LUTZ, R. W. and BÜHLMANN, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statist. Sinica* **16** 471–494. MR2267246

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, *Vol.* 1 50–55. Amer. Statist. Assoc., Alexandria, VA.

MAIER, E.-M., STÖCKER, A., FITZENBERGER, B. and GREVEN, S. (2025). Supplement to "Additive Density-on-Scalar Regression in Bayes Hilbert Spaces with an Application to Gender Economics." https://doi.org/10.1214/24-AOAS1979SUPPA, https://doi.org/10.1214/24-AOAS1979SUPPB

MARRON, J. S. and DRYDEN, I. L. (2021). *Object Oriented Data Analysis*. CRC Press, Boca Raton.

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

PARK, J. Y. and QIAN, J. (2012). Functional regression of continuous state distributions. *J. Econometrics* **167** 397–412. MR2892084 https://doi.org/10.1016/j.jeconom.2011.09.024

PAWLOWSKY-GLAHN, V., EGOZCUE, J. J. and TOLOSANA-DELGADO, R. (2015). *Modeling and Analysis of Compositional Data*. *Statistics in Practice*. Wiley, Chichester. MR3328965

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc.*, *Ser. C* **54** 507–554. MR2137253 https://doi.org/10.1111/j.1467-9876.2005.00510.x

RÜGAMER, D. and GREVEN, S. (2020). Inference for L_2-boosting. *Stat. Comput.* **30** 279–289. MR4064622 https://doi.org/10.1007/s11222-019-09882-0

SPRENGHOLZ, M., WIEBER, A. and HOLST, E. (2022). Gender identity and wives' labor market outcomes in West and East Germany between 1983 and 2016. *Socio-Econ*. *Rev*. **20** 257–279.

SRIVASTAVA, A., JERMYN, I. and JOSHI, S. (2007). Riemannian analysis of probability density functions with applications in vision. In 2007 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8.

STÖCKER, A., BROCKHAUS, S., SCHAFFER, S. A., VON BRONK, B., OPITZ, M. and GREVEN, S. (2021). Boosting functional response models for location, scale and shape with an application to bacterial competition. *Stat*. *Model*. **21** 385–404. MR4307776 https://doi.org/10.1177/1471082X20917586

TALSKÁ, R., MENAFOGLIO, A., MACHALOVÁ, J., HRON, K. and FIŠEROVÁ, E. (2018). Compositional regression with functional response. *Comput*. *Statist*. *Data Anal*. **123** 66–85. MR3777086 https://doi.org/10.1016/j.csda.2018.01.018

VAN DEN BOOGAART, K. G., EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2010). Bayes linear spaces. *SORT* **34** 201–221. MR2797479

VAN DEN BOOGAART, K. G., EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2014). Bayes Hilbert spaces. *Aust*. *N. Z. J. Stat*. **56** 171–194. MR3226435 https://doi.org/10.1111/anzs.12074

WOOD, S. N. (2017). *Generalized Additive Models*: *An Introduction with R*, 2nd ed. CRC Press, Boca Raton, FL.

YANG, H., BALADANDAYUTHAPANI, V., RAO, A. U. K. and MORRIS, J. S. (2018). Regression analyses of distributions using quantile functional regression. arXiv preprint. Available at arXiv:1810.03496.

ZHAO, Y., DATTA, A., TANG, B., ZIPUNNIKOV, V. and CAFFO, B. S. (2023). Density-on-density regression. arXiv preprint. Available at arXiv:2307.03642.

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 701–719
https://doi.org/10.1214/24-AOAS1983
This research was funded, in whole or in part, by [National Natural Science Foundation of China, 72173140, 71991474].

ENDOGENEITY AND MOMENTS IN TIME SERIES MOMENTUM'S PREDICTABILITY TEST

BY LEI JIANG[1,a], LIANG PENG[2,b], ZHONGLING QIN[3,c] AND BINGDUO YANG[4,d]

[1]*Ambassador Crawford College of Business and Entrepreneurship, Kent State University,* [a]*ljiang10@kent.edu*

[2]*Maurice R. Greenberg School of Risk Science, Robinson College of Business, Georgia State University,* [b]*lpeng@gsu.edu*

[3]*Department of Finance, Harbert College of Business, Auburn University,* [c]*zzq0018@auburn.edu*

[4]*School of Finance, Guangdong University of Finance and Economics,* [d]*bdyang2006@sina.com*

The predictability test is critical in developing momentum trading strategies. In time-series momentum (TSM) tests via linear predictive regressions, the classical and Newey–West t-tests have size distortions because of endogeneity and a lack of enough finite moments. To tackle these issues, this paper proposes a new test that features a model of the error correlations, weighted least squares estimation, and the random weighted bootstrap method. Simulations confirm its accurate size and good power. Empirically, we revisit the evidence in (*J. Financ. Econ.* **135** (2020) 774–794) and find that the TSM predictability detected by the new test is much more widespread than that by the Newey–West t-tests.

REFERENCES

BALI, T. G., CAKICI, N. and WHITELAW, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *J. Financ. Econ.* **99** 427–446.

BASRAK, B., DAVIS, R. A. and MIKOSCH, T. (2002). Regular variation of GARCH processes. *Stochastic Process. Appl.* **99** 95–115. MR1894253 https://doi.org/10.1016/S0304-4149(01)00156-9

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051 https://doi.org/10.1016/0304-4076(86)90063-1

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1007. MR0666121 https://doi.org/10.2307/1912773

GEORGOPOULOU, A. and WANG, J. (2017). The trend is your friend: Time series momentum strategies across equity and commodity markets. *Rev. Finance* **21** 1557–1592.

GOULDING, C. L., HARVEY, C. R. and MAZZOLENI, M. G. (2023). Momentum turning points. *J. Financ. Econ.* **149** 378–406.

GOYAL, A. and JEGADEESH, N. (2018). Cross-sectional and time-series tests of return predictability: What is the difference? *Rev. Financ. Stud.* **31** 1784–1824.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application. Probability and Mathematical Statistics*. Academic Press, New York. MR0624435

HARVEY, C. R. and SIDDIQUE, A. (2000). Conditional skewness in asset pricing tests. *J. Finance* **55** 1263–1295.

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204

HSING, T. (1991). On tail index estimation using dependent data. *Ann. Statist.* **19** 1547–1569. MR1126337 https://doi.org/10.1214/aos/1176348261

HUANG, D., LI, J., WANG, L. and ZHOU, G. (2020). Time series momentum: Is it there? *J. Financ. Econ.* **135** 774–794.

HURST, B., OOI, Y. H. and PEDERSEN, L. H. (2017). A century of evidence on trend-following investing. *J. Portf. Manag.* **44** 15–29.

JEGADEESH, N. and TITMAN, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *J. Finance* **48** 65–91.

JIN, Z., YING, Z. and WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88** 381–390. MR1844838 https://doi.org/10.1093/biomet/88.2.381

Key words and phrases. Endogeneity, time series momentum, predictability test, predictive regression, random weighted bootstrap.

KIM, A. Y., TSE, Y. and WALD, J. K. (2016). Time series momentum and volatility scaling. *J. Financ. Mark.* **30** 103–124.

LAZARUS, E., LEWIS, D. J., STOCK, J. H. and WATSON, M. W. (2018). HAR inference: Recommendations for practice. *J. Bus. Econom. Statist.* **36** 541–559. MR3871697 https://doi.org/10.1080/07350015.2018.1506926

LING, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/IGARCH models. *J. Econometrics* **140** 849–873. MR2408929 https://doi.org/10.1016/j.jeconom.2006.07.016

LIU, J. M., CHEN, R. and YAO, Q. (2010). Nonparametric transfer function models. *J. Econometrics* **157** 151–164. MR2652287 https://doi.org/10.1016/j.jeconom.2009.10.029

MOSKOWITZ, T., OOI, Y. and PEDERSEN, L. (2012). Time series momentum. *J. Financ. Econ.* **104** 228–250.

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864 https://doi.org/10.2307/1913610

XIAO, Z., LINTON, O. B., CARROLL, R. J. and MAMMEN, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *J. Amer. Statist. Assoc.* **98** 980–992. MR2041486 https://doi.org/10.1198/016214503000000936

ZHANG, R. and LING, S. (2015). Asymptotic inference for AR models with heavy-tailed G-GARCH noises. *Econometric Theory* **31** 880–890. MR3377272 https://doi.org/10.1017/S0266466614000632

ZHU, K. (2016). Bootstrapping the portmanteau tests in weak auto-regressive moving average models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 463–485. MR3454205 https://doi.org/10.1111/rssb.12112

ZHU, K. and LING, S. (2011). Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA-GARCH/IGARCH models. *Ann. Statist.* **39** 2131–2163. MR2893864 https://doi.org/10.1214/11-AOS895

ZHU, K. and LING, S. (2015). LADE-based inference for ARMA models with unspecified and heavy-tailed heteroscedastic noises. *J. Amer. Statist. Assoc.* **110** 784–794. MR3367264 https://doi.org/10.1080/01621459.2014.977386

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 720–743
https://doi.org/10.1214/24-AOAS1985

INFERRING MECHANISTIC PARAMETERS OF SOMATIC HYPERMUTATION USING NEURAL NETWORKS AND APPROXIMATE BAYESIAN COMPUTATION

BY THAYER FISHER[1,a], KEVIN SUNG[2,c], NOAH SIMON[1,b], JULIA FUKUYAMA[3,e] AND FREDERICK A. MATSEN IV[2,d]

[1]*Department of Biostatistics, University of Washington,* [a]*thayerf@uw.edu,* [b]*nrsimon@uw.edu*
[2]*Computational Biology Program, Fred Hutchinson Cancer Center,* [c]*ksung2@fredhutch.org,* [d]*matsen@fredhutch.org*
[3]*Department of Statistics, Indiana University,* [e]*jfukuyam@indiana.edu*

Somatic hypermutation (SHM) is a critical enzyme-mediated process of the adaptive immune response in which antibodies acquire mutations to enhance antigen binding. Despite abundant research elucidating the biochemical basis of SHM, and substantial sequence data available for parameterization, previous computational models of SHM have not been explicitly mechanistic. In this paper we bridge this gap by developing a probabilistic latent variable model encapsulating a sequence of interacting steps, thus formulating the biochemical underpinnings of SHM in a mathematical framework. However, fitting this latent variable model is challenging. To navigate this complexity, we employ an approximate Bayesian computation strategy integrated with neural networks. We are able to estimate almost all of the parameters of the model to good accuracy but find that the parameters involving the boundaries of the nucleotide stripping process are slightly more challenging, given the type of data available.

REFERENCES

BONSIGNORI, M., KREIDER, E. F., FERA, D., MEYERHOFF, R. R., BRADLEY, T., WIEHE, K., ALAM, S. M., AUSSEDAT, B., WALKOWICZ, W. E. et al. (2017). Staged induction of HIV-1 glycan-dependent broadly neutralizing antibodies. *Sci. Transl. Med.* **9**. https://doi.org/10.1126/scitranslmed.aai7514

BRINEY, B., INDERBITZIN, A., JOYCE, C. and BURTON, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566** 393–397. https://doi.org/10.1038/s41586-019-0879-y

BRINEY, B. S., WILLIS, J. R. and CROWE, J. E. JR (2012). Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun.* **13** 523–529. https://doi.org/10.1038/gene.2012.28

CHELICO, L., PHAM, P., CALABRESE, P. and GOODMAN, M. F. (2006). APOBEC3G DNA deaminase acts processively $3' \to 5'$ on single-stranded DNA. *Nat. Struct. Mol. Biol.* **13** 392–399. https://doi.org/10.1038/nsmb1086

CUI, A., DI NIRO, R., VANDER HEIDEN, J. A., BRIGGS, A. W., ADAMS, K., GILBERT, T., O'CONNOR, K. C., VIGNEAULT, F., SHLOMCHIK, M. J. et al. (2016). A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J. Immunol.* **197** 3566–3574. https://doi.org/10.4049/jimmunol.1502263

DHAR, A., DAVIDSEN, K., MATSEN, F. A. 4TH and MININ, V. N. (2018). Predicting B cell receptor substitution profiles using public repertoire data. *PLoS Comput. Biol.* **14** e1006388. https://doi.org/10.1371/journal.pcbi.1006388

DUNN-WALTERS, D. K., DOGAN, A., BOURSIER, L., MACDONALD, C. M. and SPENCER, J. (1998). Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *J. Immunol.* **160** 2360–2364.

DUNN-WALTERS, D. K. and SPENCER, J. (1998). Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology* **95** 339–345.

FEARNHEAD, P. and PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 419–474. MR2925370 https://doi.org/10.1111/j.1467-9868.2011.01010.x

Key words and phrases. Bayesian statistics, likelihood-free inference, computational biology, adaptive immunology.

FENG, J., SHAW, D. A., MININ, V. N., SIMON, N. and MATSEN, F. A. IV (2019). Survival analysis of DNA mutation motifs with penalized proportional hazards. *Ann. Appl. Stat.* **13** 1268–1294. MR3963571 https://doi.org/10.1214/18-AOAS1233

FISHER, T., LUEDTKE, A., CARONE, M. and SIMON, N. (2023). Deep learning for marginal Bayesian posterior inference with recurrent neural networks. https://doi.org/10.5705/ss.202020.0348

FISHER, T., SUNG, K., SIMON, N., FUKUYAMA, J. and MATSEN IV, F. A. (2025). Supplement to "Inferring mechanistic parameters of somatic hypermutation using neural networks and approximate Bayesian computation." https://doi.org/10.1214/24-AOAS1985SUPPA, https://doi.org/10.1214/24-AOAS1985SUPPB

FLAGEL, L., BRANDVAIN, Y. and SCHRIDER, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol. Biol. Evol.* **36** 220–238. https://doi.org/10.1093/molbev/msy224

GREEN, B., BELCHEVA, A., NEPAL, R. M., BOULIANNE, B. and MARTIN, A. (2011). The mismatch repair pathway functions normally at a non-AID target in germinal center B cells. *Blood* **118** 3013–3018. https://doi.org/10.1182/blood-2011-03-345991

HERSHBERG, U., UDUMAN, M., SHLOMCHIK, M. J. and KLEINSTEIN, S. H. (2008). Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.* **20** 683–694. https://doi.org/10.1093/intimm/dxn026

HWANG, J. K., WANG, C., DU, Z., MEYERS, R. M., KEPLER, T. B., NEUBERG, D., KWONG, P. D., MASCOLA, J. R., JOYCE, M. G. et al. (2017). Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc. Natl. Acad. Sci. USA* **114** 8614–8619. https://doi.org/10.1073/pnas.1709203114

JANSEN, J. G., LANGERAK, P., TSAALBI-SHTYLIK, A., VAN DEN BERK, P., JACOBS, H. and DE WIND, N. (2006). Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. *J. Exp. Med.* **203** 319–323.

JARDINE, J., JULIEN, J.-P., MENIS, S., OTA, T., KALYUZHNIY, O., MCGUIRE, A., SOK, D., HUANG, P.-S., MACPHERSON, S. et al. (2013). Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340** 711–716. https://doi.org/10.1126/science.1234150

JIANG, B., WU, T.-Y., ZHENG, C. and WONG, W. H. (2017). Learning summary statistic for approximate Bayesian computation via deep neural network. *Statist. Sinica* **27** 1595–1618. MR3701500

KADYROV, F. A., DZANTIEV, L., CONSTANTIN, N. and MODRICH, P. (2006). Endonucleolytic function of MutLa in human mismatch repair. *Cell* **126** 297–308. https://doi.org/10.1016/j.cell.2006.05.039

KIMURA, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16** 111–120. https://doi.org/10.1007/BF01731581

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KLEINSTEIN, S. H., LOUZOUN, Y. and SHLOMCHIK, M. J. (2003). Estimating hypermutation rates from clonal tree data. *J. Immunol.* **171** 4639–4649.

KRANTSEVICH, A., TANG, C. and MACCARTHY, T. (2020). Correlations in somatic hypermutation between sites in IGHV genes can be explained by interactions between AID and/or Polη hotspots. *Front. Immunol.* **11** 618409. https://doi.org/10.3389/fimmu.2020.618409

KRIJGER, P. H. L., LANGERAK, P., VAN DEN BERK, P. C. M. and JACOBS, H. (2009). Dependence of nucleotide substitutions on Ung2, Msh2, and PCNA-Ub during somatic hypermutation. *J. Exp. Med.* **206** 2603–2611. https://doi.org/10.1084/jem.20091707

LANGERAK, P., NYGREN, A. O., KRIJGER, P. H., VAN DEN BERK, P. C. and JACOBS, H. (2007). A/T mutagenesis in hypermutated immunoglobulin genes strongly depends on PCNAK164 modification. *J. Exp. Med.* **204** 1989–1998.

LUPO, C., SPISAK, N., WALCZAK, A. M. and MORA, T. (2022). Learning the statistics and landscape of somatic mutation-induced insertions and deletions in antibodies. *PLoS Comput. Biol.* **18** e1010167. https://doi.org/10.1371/journal.pcbi.1010167

MAK, C. H., PHAM, P., AFIF, S. A. and GOODMAN, M. F. (2013). A mathematical model for scanning and catalysis on single-stranded DNA, illustrated with activation-induced deoxycytidine deaminase. *J. Biol. Chem.* **288** 29786–29795. https://doi.org/10.1074/jbc.M113.506550

MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.

MCCOY, C. O., BEDFORD, T., MININ, V. N., BRADLEY, P., ROBINS, H. and MATSEN IV, F. A. (2015). Quantifying evolutionary constraints on B-cell affinity maturation. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370** 20140244. https://doi.org/10.1098/rstb.2014.0244

METHOT, S. P. and DI NOIA, J. M. (2017). Chapter two—Molecular mechanisms of somatic hypermutation and class switch recombination. In *Advances in Immunology* (F. W. Alt, ed.) **133** 37–87. Academic Press, San Diego.

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

NEUBERGER, M. S., DI NOIA, J. M., BEALE, R. C. L., WILLIAMS, G. T., YANG, Z. and RADA, C. (2005). Somatic hypermutation at A:T pairs: Polymerase error versus dUTP incorporation. *Nat. Rev., Immunol.* **5** 171–178.

NGUYEN, L.-T., SCHMIDT, H. A., VON HAESELER, A. and MINH, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32** 268–274. https://doi.org/10.1093/molbev/msu300

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**.

PHAM, P., BRANSTEITTER, R., PETRUSKA, J. and GOODMAN, M. F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424** 103–107. https://doi.org/10.1038/nature01760

PHAM, P., MALIK, S., MAK, C., CALABRESE, P. C., ROEDER, R. G. and GOODMAN, M. F. (2019). AID-RNA polymerase II transcription-dependent deamination of IgV DNA. *Nucleic Acids Res.* **47** 10815–10829. https://doi.org/10.1093/nar/gkz821

PILZECKER, B. and JACOBS, H. (2019). Mutating for good: DNA damage responses during somatic hypermutation. *Front. Immunol.* **10** 438. https://doi.org/10.3389/fimmu.2019.00438

ROGOZIN, I. B. and DIAZ, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J. Immunol.* **172** 3382–3384.

ROGOZIN, I. B., PAVLOV, Y. I., BEBENEK, K., MATSUDA, T. and KUNKEL, T. A. (2001). Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat. Immunol.* **2** 530–536. https://doi.org/10.1038/88732

RUSSELL, M. L., SIMON, N., BRADLEY, P. and MATSEN, F. A. IV (2023). Statistical inference reveals the role of length, GC content, and local sequence in V(D)J nucleotide trimming. *eLife* **12** e85145. https://doi.org/10.7554/eLife.85145

SARIBASAK, H., MAUL, R. W., CAO, Z., YANG, W. W., SCHENTEN, D., KRACKER, S. and GEARHART, P. J. (2012). DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J. Exp. Med.* **209** 1075–1081.

SENAVIRATHNE, G., BERTRAM, J. G., JASZCZUR, M., CHAURASIYA, K. R., PHAM, P., MAK, C. H., GOODMAN, M. F. and RUEDA, D. (2015). Activation-induced deoxycytidine deaminase (AID) co-transcriptional scanning at single-molecule resolution. *Nat. Commun.* **6** 10209.

SHEEHAN, S. and SONG, Y. S. (2016). Deep learning for population genetic inference. *PLoS Comput. Biol.* **12** e1004845. https://doi.org/10.1371/journal.pcbi.1004845

SHEN, H. M., TANAKA, A., BOZEK, G., NICOLAE, D. and STORB, U. (2006). Somatic hypermutation and class switch recombination in Msh6-/- Ung-/- double-knockout mice. *J. Immunol.* **177** 5386–5392.

SHENG, Z., SCHRAMM, C. A., CONNORS, M., MORRIS, L., MASCOLA, J. R., KWONG, P. D. and SHAPIRO, L. (2016). Effects of Darwinian selection and mutability on rate of broadly neutralizing antibody evolution during HIV-1 infection. *PLoS Comput. Biol.* **12** e1004940. https://doi.org/10.1371/journal.pcbi.1004940

SPENCER, J., DUNN, M. and DUNN-WALTERS, D. K. (1999). Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. *J. Immunol.* **162** 6596–6601.

SPISAK, N., WALCZAK, A. M. and MORA, T. (2020). Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Res.* **48** 10702–10712. https://doi.org/10.1093/nar/gkaa825

SVOZIL, D., KVASNICKA, V. and POSPICHAL, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **39** 43–62.

TANG, C., KRANTSEVICH, A. and MACCARTHY, T. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience* **25** 103668. https://doi.org/10.1016/j.isci.2021.103668

UDUMAN, M., YAARI, G., HERSHBERG, U., STERN, J. A., SHLOMCHIK, M. J. and KLEINSTEIN, S. H. (2011). Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* **39** W499–W504. https://doi.org/10.1093/nar/gkr413

UNNIRAMAN, S. and SCHATZ, D. G. (2007). Strand-biased spreading of mutations during somatic hypermutation. *Science* **317** 1227–1230. https://doi.org/10.1126/science.1145065

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. and POLOSUKHIN, I. (2017). Attention is all you need.

VIEIRA, M. C., ZINDER, D. and COBEY, S. (2018). Selection and neutral mutations drive pervasive mutability losses in long-lived anti-HIV B-cell lineages. *Mol. Biol. Evol.* **35** 1135–1146. https://doi.org/10.1093/molbev/msy024

WANG, M., RADA, C. and NEUBERGER, M. S. (2010). Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *J. Exp. Med.* **207** 141–153. https://doi.org/10.1084/jem.20092238

WANG, Y., ZHANG, S., YANG, X., HWANG, J. K., ZHAN, C., LIAN, C., WANG, C., GUI, T., WANG, B. et al. (2023). Mesoscale DNA feature in antibody-coding sequence facilitates somatic hypermutation. *Cell* **186** 2193–2207. https://doi.org/10.1016/j.cell.2023.03.030

WEI, L., CHAHWAN, R., WANG, S., WANG, X., PHAM, P. T., GOODMAN, M. F., BERGMAN, A., SCHARFF, M. D. and MACCARTHY, T. (2015). Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. USA* **112** E728–E737. https://doi.org/10.1073/pnas.1500788112

WIEHE, K., BRADLEY, T., RYAN MEYERHOFF, R., HART, C., WILLIAMS, W. B., EASTERHOFF, D., FAISON, W. J., KEPLER, T. B., SAUNDERS, K. O. et al. (2018). Functional relevance of improbable antibody mutations for HIV broadly neutralizing antibody development. *Cell Host Microbe* **23** 759–765. https://doi.org/10.1016/j.chom.2018.04.018

WILSON, T. M., VAISMAN, A., MARTOMO, S. A., SULLIVAN, P., LAN, L., HANAOKA, F., YASUI, A., WOODGATE, R. and GEARHART, P. J. (2005). MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. *J. Exp. Med.* **201** 637–645.

YAARI, G., BENICHOU, J. I. C., HEIDEN, J. A. V., KLEINSTEIN, S. H. and LOUZOUN, Y. (2015). The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370**. https://doi.org/10.1098/rstb.2014.0242

YAARI, G., UDUMAN, M. and KLEINSTEIN, S. H. (2012). Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40** e134. https://doi.org/10.1093/nar/gks457

YAARI, G., VANDER HEIDEN, J. A., UDUMAN, M., GADALA-MARIA, D., GUPTA, N., STERN, J. N. H., O'CONNOR, K. C., HAFLER, D. A., LASERSON, U. et al. (2013). Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* **4** 358. https://doi.org/10.3389/fimmu.2013.00358

YEAP, L.-S., HWANG, J. K., DU, Z., MEYERS, R. M., MENG, F.-L., JAKUBAUSKAITE, A., LIU, M., MANI, V., NEUBERG, D. et al. (2015). Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* **163** 1124–1137. https://doi.org/10.1016/j.cell.2015.10.042

ZIVOJNOVIC, M., DELBOS, F., ZUBANI, G. G., JULÉ, A., ALCAIS, A., WEILL, J.-C., REYNAUD, C.-A. and STORCK, S. (2014). Somatic hypermutation at A/T-rich oligonucleotide substrates shows different strand polarities in Ung-deficient or -proficient backgrounds. *Mol. Cell. Biol.* **34** 2176–2187. https://doi.org/10.1128/MCB.01452-13

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 744–768
https://doi.org/10.1214/24-AOAS1986

DESIGN-BASED INFERENCE FOR SPATIAL EXPERIMENTS UNDER UNKNOWN INTERFERENCE

BY YE WANG[1,a], CYRUS SAMII[2,b], HAOGE CHANG[3,c] AND PM ARONOW[4,d]

[1]*Department of Political Science, University of North Carolina, Chapel Hill,* [a]*yewang@unc.edu*
[2]*Politics Department, New York University,* [b]*cds2083@nyu.edu*
[3]*Department of Economics, Columbia University,* [c]*hc3615@columbia.edu*
[4]*Departments of Statistics and Data Science, Political Science, Biostatistics and Economics, Yale University,* [d]*p.aronow@yale.edu*

We consider design-based causal inference for spatial experiments in which treatments may have effects that bleed out and feed back in complex ways. Such spatial spillover effects violate the "no interference" assumption for standard causal inference methods. The complexity of spatial spillover effects also raises the risk of misspecification and bias in model-based analyses. We offer an approach for robust inference in such settings without having to specify a parametric outcome model. We define a spatial "average marginalized effect" (AME) that characterizes how, in expectation, units of observation that are a specified distance from an intervention location are affected by treatment at that location, averaging over effects emanating from other intervention nodes. We show that randomization is sufficient for nonparametric identification of the AME, even if the nature of interference is unknown. Under mild restrictions on the extent of interference, we establish asymptotic distributions of estimators and provide methods for both sample-theoretic and randomization-based inference. We show conditions under which the AME recovers a structural effect. We illustrate our approach with a simulation study. Then we reanalyze a randomized field experiment and a quasi-experiment on forest conservation, showing how our approach offers robust inference on policy-relevant spillover effects.

REFERENCES

ALIX-GARCIA, J. M., SHAPIRO, E. N. and SIMS, K. R. (2012). Forest conservation and slippage: Evidence from Mexico's national payments for ecosystem services program. *Land Econ*. **88** 613–638.

ANDREWS, D. W. K. (1984). Nonstrong mixing autoregressive processes. *J. Appl. Probab.* **21** 930–934. MR0766830 https://doi.org/10.2307/3213710

ARBIA, G. (2006). *Spatial Econometrics*: *Statistical Foundations and Applications to Regional Convergence*. *Advances in Spatial Science*. Springer, Berlin. MR2221789

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

ASQUITH, N. (2020). Large-scale randomized control trials of incentive-based conservation: What have we learned? *World Dev.* **127** 104785.

BELL, R. M. and MCCAFFREY, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Surv. Methodol.* **28** 169–181.

CHANG, H. (2023). Design-based estimation theory for complex experiments. arXiv preprint. Available at arXiv:2311.06891.

CONLEY, T. G. (1999). GMM estimation with cross sectional dependence. *J. Econometrics* **92** 1–45. MR1707000 https://doi.org/10.1016/S0304-4076(98)00084-0

COX, D. R. (1958). *Planning of Experiments*. *A Wiley Publication in Applied Statistics*. Wiley, New York. MR0095561

DARMOFAL, D. (2015). *Spatial Analysis for the Social Sciences*. Cambridge Univ. Press, Cambridge.

Key words and phrases. Causal inference, design-based inference, interference.

DOUKHAN, P. and LANG, G. (2002). Rates in the empirical central limit theorem for stationary weakly dependent random fields. *Stat. Inference Stoch. Process.* **5** 199–228. MR1917292 https://doi.org/10.1023/A:1016396906927

FERRARO, P. J., HANAUER, M. M. and SIMS, K. R. (2011). Conditions associated with protected area success in conservation and poverty reduction. *Proc. Natl. Acad. Sci. USA* **108** 13913–13918.

GAO, M. and DING, P. (2023). Causal inference in network experiments: regression-based analysis and design-based properties. arXiv preprint. Available at arXiv:2309.07476.

GIBBONS, S. and OVERMAN, H. G. (2012). Mostly pointless spatial econometrics? *J. Reg. Sci.* **52** 172–191.

GOLGHER, A. B. and VOSS, P. R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spat. Demogr.* **4** 175–205.

HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference in infectious diseases. *Epidemiology* **6** 142–151. https://doi.org/10.1097/00001648-199503000-00010

HANSEN, M. C., POTAPOV, P. V., MOORE, R., HANCHER, M., TURUBANOVA, S. A., TYUKAVINA, A., THAU, D., STEHMAN, S. V., GOETZ, S. J. et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science* **342** 850–853.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

HORN, R. A. and JOHNSON, C. R. (2013). *Matrix Analysis*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2978290

HU, Y., LI, S. and WAGER, S. (2022). Average direct and indirect causal effects under interference. *Biometrika* **109** 1165–1172. MR4519123 https://doi.org/10.1093/biomet/asac008

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 KOLESAR, M. (2012). Robust Standard Errors in Small Samples: Some Practical Advice. NBER Working Paper Series 14726.

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

JAYACHANDRAN, S., DE LAAT, J., LAMBIN, E. F., STANTON, C. Y., AUDY, R. and THOMAS, N. E. (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science* **357** 267–273. https://doi.org/10.1126/science.aan0568

JENISH, N. (2016). Spatial semiparametric model with endogenous regressors. *Econometric Theory* **32** 714–739. MR3506437 https://doi.org/10.1017/S0266466614000905

JENISH, N. and PRUCHA, I. R. (2009). Central limit theorems and uniform laws of large numbers for arrays of random fields. *J. Econometrics* **150** 86–98. MR2525996 https://doi.org/10.1016/j.jeconom.2009.02.009

JENISH, N. and PRUCHA, I. R. (2012). On spatial processes and asymptotic inference under near-epoch dependence. *J. Econometrics* **170** 178–190. MR2955948 https://doi.org/10.1016/j.jeconom.2012.05.022

KELEJIAN, H. and PIRAS, G. (2017). *Spatial Econometrics*. Elsevier, New York.

KELEJIAN, H. H. and PRUCHA, I. R. (2007). HAC estimation in a spatial framework. *J. Econometrics* **140** 131–154. MR2395919 https://doi.org/10.1016/j.jeconom.2006.09.005

LEUNG, M. P. (2022). Rate-optimal cluster-randomized designs for spatial interference. *Ann. Statist.* **50** 3064–3087. MR4500634 https://doi.org/10.1214/22-aos2224

LEUNG, M. P. (2023). Design of cluster-randomized trials with cross-cluster interference. arXiv preprint. Available at arXiv:2310.18836.

LI, S. and WAGER, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *Ann. Statist.* **50** 2334–2358. MR4474493 https://doi.org/10.1214/22-aos2191

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864 https://doi.org/10.2307/1913610

OGBURN, E. L., SOFRYGIN, O., DÍAZ, I. and VAN DER LAAN, M. J. (2024). Causal inference for social network data. *J. Amer. Statist. Assoc.* **119** 597–611. MR4713916 https://doi.org/10.1080/01621459.2022.2131557

PAPADOGEORGOU, G., IMAI, K., LYALL, J. and LI, F. (2022). Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in Iraq. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1969–1999. MR4515563 https://doi.org/10.1111/rssb.12548

REICH, B. J., YANG, S., GUAN, Y., GIFFIN, A. B., MILLER, M. J. and RAPPOLD, A. (2021). A review of spatial causal inference methods for environmental and epidemiological applications. *Int. Stat. Rev.* **89** 605–634. MR4411920 https://doi.org/10.1111/insr.12452

SAMII, C., LISIECKI, M., KULKARNI, P., PALER, L., CHAVIS, L., SNILSTVEIT, B., VOJTKOVA, M. and GALLAGHER, E. (2014). Effects of payment for environmental services (PES) on deforestation and poverty in low and middle income countries: A systematic review. *Campbell Syst. Rev.* **10** 1–95.

SÄVJE, F., ARONOW, P. M. and HUDGENS, M. G. (2021). Average treatment effects in the presence of unknown interference. *Ann. Statist.* **49** 673–701. MR4255103 https://doi.org/10.1214/20-aos1973

VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2011). Effect partitioning under interference in two-stage randomized vaccine trials. *Statist. Probab. Lett.* **81** 861–869. MR2793754 https://doi.org/10.1016/j.spl.2011.02.019

WANG, Y., SAMII, C., CHANG, H. and ARONOW, P. (2025). Supplement to "Design-based inference for spatial experiments under unknown interference." https://doi.org/10.1214/24-AOAS1986SUPPA, https://doi.org/10.1214/24-AOAS1986SUPPB

WUNDER, S. (2008). How do we deal with leakage. In *Moving Ahead with REDD: Issues, Options and Implications* **1** 65–75.

YOUNG, A. (2015). Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections. London School of Economics. Unpublished Manuscript.

ZIGLER, C. M. and PAPADOGEORGOU, G. (2021). Bipartite causal inference with interference. *Statist. Sci.* **36** 109–123. MR4194206 https://doi.org/10.1214/19-STS749

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 769–797
https://doi.org/10.1214/24-AOAS1988

LOW-RANK LONGITUDINAL FACTOR REGRESSION WITH APPLICATION TO CHEMICAL MIXTURES

BY GLENN PALMER[a], AMY H. HERRING[b] AND DAVID B. DUNSON[c]

Department of Statistical Science, Duke University, [a]*glenn.palmer@duke.edu,* [b]*amy.herring@duke.edu,* [c]*dunson@duke.edu*

Developmental epidemiology commonly focuses on assessing the association between multiple early life exposures and childhood health. Statistical analyses of data from such studies focus on inferring the contributions of individual exposures, while also characterizing time-varying and interacting effects. Such inferences are made more challenging by correlations among exposures, nonlinearity, and the curse of dimensionality. Motivated by studying the effects of prenatal bisphenol A (BPA) and phthalate exposures on glucose metabolism in adolescence using data from the ELEMENT study, we propose a low-rank longitudinal factor regression (LowFR) model for tractable inference on flexible longitudinal exposure effects. LowFR handles highly-correlated exposures using a Bayesian dynamic factor model, which is fit jointly with a health outcome via a novel factor regression approach. The model collapses on simpler and intuitive submodels when appropriate, while expanding to allow considerable flexibility in time-varying and interaction effects when supported by the data. After demonstrating LowFR's effectiveness in simulations, we use it to analyze the ELEMENT data and find that diethyl and dibutyl phthalate metabolite levels in trimesters 1 and 2 are associated with altered glucose metabolism in adolescence.

REFERENCES

ALONSO-MAGDALENA, P., MORIMOTO, S., RIPOLL, C., FUENTES, E. and NADAL, A. (2006). The estrogenic effect of bisphenol A disrupts pancreatic beta-cell function in vivo and induces insulin resistance. *Environ. Health Perspect.* **114** 106–112. https://doi.org/10.1289/ehp.8451

ANTONELLI, J., WILSON, A. and COULL, B. A. (2024). Multiple exposure distributed lag models with variable selection. *Biostatistics* **25** 1–19. MR4678528 https://doi.org/10.1093/biostatistics/kxac038

BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2849614 https://doi.org/10.1002/9781119970583

BECKER, K., GÜEN, T., SEIWERT, M., CONRAD, A., PICK-FUSS, H., MÜLLER, J., WITTASSEK, M., SCHULZ, C. and KOLOSSA-GEHRING, M. (2009). GerES IV: Phthalate metabolites and bisphenol A in urine of German children. *Int. J. Hyg. Environ. Health* **212** 685–692.

BELLO, G. A., ARORA, M., AUSTIN, C., HORTON, M. K., WRIGHT, R. O. and GENNINGS, C. (2017). Extending the distributed lag model framework to handle chemical mixtures. *Environ. Res.* **156** 253–264.

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

BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

BOBB, J. F., CLAUS HENN, B., VALERI, L. and COULL, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ. Health* **17** 1–10.

BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. MR3365442 https://doi.org/10.1093/biostatistics/kxu058

BORNEHAG, C.-G., SUNDELL, J., WESCHLER, C. J., SIGSGAARD, T., LUNDGREN, B., HASSELGREN, M. and HÄGERHED-ENGMAN, L. (2004). The association between asthma and allergic symptoms in children and phthalates in house dust: A nested case-control study. *Environ. Health Perspect.* **112** 1393–1397. https://doi.org/10.1289/ehp.7187

Key words and phrases. Factor analysis, Bayesian statistics, interaction effects, longitudinal data analysis, mixtures problem, maternal and child health.

BRAUN, J. M., SATHYANARAYANA, S. and HAUSER, R. (2013). Phthalate exposure and children's health. *Curr. Opin. Pediatr.* **25** 247.

BUCKLEY, J. P., HAMRA, G. B. and BRAUN, J. M. (2019). Statistical approaches for investigating periods of susceptibility in children's environmental health research. *Curr. Environ. Health Rep.* **6** 1–7. https://doi.org/10.1007/s40572-019-0224-5

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**.

CARRICO, C., GENNINGS, C., WHEELER, D. C. and FACTOR-LITVAK, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J. Agric. Biol. Environ. Stat.* **20** 100–120. MR3334469 https://doi.org/10.1007/s13253-014-0180-3

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* 73–80. PMLR.

CHANG, H. H., WARREN, J. L., DARROW, L. A., REICH, B. J. and WALLER, L. A. (2015). Assessment of critical exposure and outcome windows in time-to-event analysis with application to air pollution and preterm birth study. *Biostatistics* **16** 509–521. MR3365443 https://doi.org/10.1093/biostatistics/kxu060

CHEN, Y.-H., MUKHERJEE, B. and BERROCAL, V. J. (2019). Distributed lag interaction models with two pollutants. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 79–97. MR3902983 https://doi.org/10.1111/rssc.12297

COLICINO, E., PEDRETTI, N. F., BUSGANG, S. A. and GENNINGS, C. (2020). Per- and poly-fluoroalkyl substances and bone mineral density: Results from the Bayesian weighted quantile sum regression. *Environ. Epidemiol.* **4** e092. https://doi.org/10.1097/EE9.0000000000000092

CURTIN, P., KELLOGG, J., CECH, N. and GENNINGS, C. (2021). A random subset implementation of weighted quantile sum ($\mathrm{WQS_{RS}}$) regression for analysis of high-dimensional mixtures. *Comm. Statist. Simulation Comput.* **50** 1119–1134. MR4246940 https://doi.org/10.1080/03610918.2019.1577971

CZARNOTA, J., GENNINGS, C. and WHEELER, D. C. (2015). Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. *Cancer Inform.* **14** 159–171. https://doi.org/10.4137/CIN.S17295

DE MICHEAUX, P. L. (2017). CompQuadForm: Distribution function of quadratic forms in normal variables. *CRAN Repository*. R package version 1.4.3.

DESVERGNE, B., FEIGE, J. N. and CASALS-CASAS, C. (2009). PPAR-mediated activity of phthalates: A link to the obesity epidemic? *Mol. Cell. Endocrinol.* **304** 43–48.

DE WAAL, D. (1985). Matrix-valued distributions. *Encycl. Statist. Sci.*

DOMINICI, F., PENG, R. D., BARR, C. D. and BELL, M. L. (2010). Opinion: Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach. *Epidemiology* 187–194.

ENG, D. S., LEE, J. M., GEBREMARIAM, A., MEEKER, J. D., PETERSON, K. and PADMANABHAN, V. (2013). Bisphenol A and chronic disease risk factors in US children. *Pediatrics* **132** e637–e645.

FERRARI, F. and DUNSON, D. B. (2021). Bayesian factor analysis for inference on interactions. *J. Amer. Statist. Assoc.* **116** 1521–1532. MR4309290 https://doi.org/10.1080/01621459.2020.1745813

FOSDICK, B. K. and HOFF, P. D. (2014). Separable factor analysis with applications to mortality data. *Ann. Appl. Stat.* **8** 120–147. MR3191985 https://doi.org/10.1214/13-AOAS694

FREDERIKSEN, H., SØRENSEN, K., MOURITSEN, A., AKSGLAEDE, L., HAGEN, C., PETERSEN, J., SKAKKEBAEK, N., ANDERSSON, A.-M. and JUUL, A. (2012). High urinary phthalate concentration associated with delayed pubarche in girls. *Int. J. Androl.* **35** 216–226.

GASPARRINI, A., ARMSTRONG, B. and KENWARD, M. G. (2010). Distributed lag non-linear models. *Stat. Med.* **29** 2224–2234. MR2757147 https://doi.org/10.1002/sim.3940

GASPARRINI, A., SCHEIPL, F., ARMSTRONG, B. and KENWARD, M. G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics* **73** 938–948. MR3713127 https://doi.org/10.1111/biom.12645

GENNINGS, C., CURTIN, P., BELLO, G., WRIGHT, R., ARORA, M. and AUSTIN, C. (2020). Lagged WQS regression for mixtures with many components. *Environ. Res.* **186** 109529. https://doi.org/10.1016/j.envres.2020.109529

GOODRICH, J. M., TANG, L., CARMONA, Y. R., MEIJER, J. L., PERNG, W., WATKINS, D. J., MEEKER, J. D., MERCADO-GARCÍA, A., CANTORAL, A. et al. (2022a). Trimester-specific phthalate exposures in pregnancy are associated with circulating metabolites in children. *PLoS ONE* **17** e0272794.

GOODRICH, J. M., TANG, L., CARMONA, Y. R., MEIJER, J. L., PERNG, W., WATKINS, D. J., MEEKER, J. D., MERCADO-GARCÍA, A., CANTORAL, A. et al. (2022b). Trimester-specific phthalate exposures in pregnancy are associated with circulating metabolites in children [data set].

GRIMALDI, M., BOULAHTOUF, A., DELFOSSE, V., THOUENNON, E., BOURGUET, W. and BALAGUER, P. (2015). Reporter cell lines for the characterization of the interactions between human nuclear receptors and endocrine disruptors. *Front. Endocrinol.* **6** 62.

GRÜN, F. and BLUMBERG, B. (2007). Perturbed nuclear receptor signaling by environmental obesogens as emerging factors in the obesity crisis. *Rev. Endocr. Metab. Disord.* **8** 161–171. https://doi.org/10.1007/s11154-007-9049-x

GÜVEN, C., DAL, F., AHBAB, M. A., TASKIN, E., AHBAB, S., CINAR, S. A., EKMEKÇI, S. S., GÜLEÇ, Ç., ABACI, N. et al. (2016). Low dose monoethyl phthalate (MEP) exposure triggers proliferation by activating PDX-1 at 1.1 B4 human pancreatic beta cells. *Food Chem. Toxicol.* **93** 41–50.

HAHN, P. R., CARVALHO, C. M. and MUKHERJEE, S. (2013). Partial factor modeling: Predictor-dependent shrinkage for linear regression. *J. Amer. Statist. Assoc.* **108** 999–1008. MR3174679 https://doi.org/10.1080/01621459.2013.779843

HAMRA, G. B. and BUCKLEY, J. P. (2018). Environmental exposure mixtures: Questions and methods to address them. *Curr. Epidemiol. Rep.* **5** 160–165. https://doi.org/10.1007/s40471-018-0145-0

HAO, N., FENG, Y. and ZHANG, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *J. Amer. Statist. Assoc.* **113** 615–625. MR3832213 https://doi.org/10.1080/01621459.2016.1264956

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

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851 https://doi.org/10.1214/07-AOAS107

HOFF, P. D. and NIU, X. (2012). A covariance regression model. *Statist. Sinica* **22** 729–753. MR2954359 https://doi.org/10.5705/ss.2010.051

HUNG, H. and WANG, C.-C. (2013). Matrix variate logistic regression model with application to EEG data. *Biostatistics* **14** 189–202.

HUNG, H., WU, P., TU, I. and HUANG, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99** 569–583. MR2966770 https://doi.org/10.1093/biomet/ass019

HURST, C. H. and WAXMAN, D. J. (2003). Activation of PPARalpha and PPARgamma by environmental phthalate monoesters. *Toxicol. Sci.* **74** 297–308. https://doi.org/10.1093/toxsci/kfg145

JAMES-TODD, T., STAHLHUT, R., MEEKER, J. D., POWELL, S.-G., HAUSER, R., HUANG, T. and RICH-EDWARDS, J. (2012). Urinary phthalate metabolite concentrations and diabetes among women in the National Health and Nutrition Examination Survey (NHANES) 2001–2008. *Environ. Health Perspect.* **120** 1307–1313.

JIANG, B., PETKOVA, E., TARPEY, T. and OGDEN, R. T. (2020). A Bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies. *Biometrics* **76** 87–97. MR4098546 https://doi.org/10.1111/biom.13151

JOUBERT, B. R., KIOUMOURTZOGLOU, M.-A., CHAMBERLAIN, T., CHEN, H. Y., GENNINGS, C., TURYK, M. E., MIRANDA, M. L., WEBSTER, T. F., ENSOR, K. B. et al. (2022). Powering research through innovative methods for mixtures in epidemiology (PRIME) program: Novel and expanded statistical methods. *Int. J. Environ. Res. Public Health* **19** 1378.

JUST, A. C., WHYATT, R. M., PERZANOWSKI, M. S., CALAFAT, A. M., PERERA, F. P., GOLDSTEIN, I. F., CHEN, Q., RUNDLE, A. G. and MILLER, R. L. (2012). Prenatal exposure to butylbenzyl phthalate and early eczema in an urban cohort. *Environ. Health Perspect.* **120** 1475–1480.

KAHN, L. G., PHILIPPAT, C., NAKAYAMA, S. F., SLAMA, R. and TRASANDE, L. (2020). Endocrine-disrupting chemicals: Implications for human health. *Lancet Diabetes Endocrinol.* **8** 703–718.

KIM, Y., HA, E.-H., KIM, E.-J., PARK, H., HA, M., KIM, J.-H., HONG, Y.-C., CHANG, N. and KIM, B.-N. (2011). Prenatal exposure to phthalates and infant development at 6 months: Prospective Mothers and Children's Environmental Health (MOCEH) study. *Environ. Health Perspect.* **119** 1495–1500.

KOCH, H., CHRISTENSEN, K., HARTH, V., LORBER, M. and BRÜNING, T. (2012). Di-n-butyl phthalate (DnBP) and diisobutyl phthalate (DiBP) metabolism in a human volunteer after single oral doses. *Arch. Toxicol.* **86** 1829–1839.

LANG, I. A., GALLOWAY, T. S., SCARLETT, A., HENLEY, W. E., DEPLEDGE, M., WALLACE, R. B. and MELZER, D. (2008). Association of urinary bisphenol A concentration with medical disorders and laboratory abnormalities in adults. *JAMA* **300** 1303–1310. https://doi.org/10.1001/jama.300.11.1303

LEGRAMANTI, S., DURANTE, D. and DUNSON, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107** 745–752. MR4138988 https://doi.org/10.1093/biomet/asaa008

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488 https://doi.org/10.2307/2669940

LIU, J. Z., DENG, W., LEE, J., LIN, P. D., VALERI, L., CHRISTIANI, D. C., BELLINGER, D. C., WRIGHT, R. O., MAZUMDAR, M. M. et al. (2022). A cross-validated ensemble approach to robust hypothesis testing of continuous nonlinear interactions: Application to nutrition-environment studies. *J. Amer. Statist. Assoc.* **117** 561–573. MR4436296 https://doi.org/10.1080/01621459.2021.1962889

LIU, S. H., BOBB, J. F., LEE, K. H. et al. (2018). Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics* **19** 325–341. MR3815175 https://doi.org/10.1093/biostatistics/kxx036

MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60** 234–256.

MORK, D. and WILSON, A. (2022). Treed distributed lag nonlinear models. *Biostatistics* **23** 754–771. MR4454953 https://doi.org/10.1093/biostatistics/kxaa051

MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Amer. Statist. Assoc.* **108** 656–665. MR3174649 https://doi.org/10.1080/01621459.2012.762328

NGUYEN, P. H., HERRING, A. H. and ENGEL, S. M. (2023). Power analysis of exposure mixture studies via Monte Carlo simulations. *Stat. Biosci.* 1–26.

OVASKAINEN, O., TIKHONOV, G., NORBERG, A., GUILLAUME BLANCHET, F., DUAN, L., DUNSON, D., ROSLIN, T. and ABREGO, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20** 561–576.

PALMER, G., HERRING, A. H and DUNSON, D. B (2025a). Supplement to "Low-rank longitudinal factor regression with application to chemical mixtures." https://doi.org/10.1214/24-AOAS1988SUPPA

PALMER, G., HERRING, A. H and DUNSON, D. B (2025b). Supplement to "Low-rank longitudinal factor regression with application to chemical mixtures." https://doi.org/10.1214/24-AOAS1988SUPPB

POWOROZNEK, E. (2020). Infinitefactor: Bayesian infinite factor models. *CRAN Repository*. R package version 1.0.

SHANIK, M. H., XU, Y., SKRHA, J., DANKNER, R., ZICK, Y. and ROTH, J. (2008). Insulin resistance and hyperinsulinemia: Is hyperinsulinemia the cart or the horse? *Diabetes Care* **31** S262–S268.

SILVA, M. J., BARR, D. B., REIDY, J. A., MALEK, N. A., HODGE, C. C., CAUDILL, S. P., BROCK, J. W., NEEDHAM, L. L. and CALAFAT, A. M. (2004). Urinary levels of seven phthalate metabolites in the U.S. population from the National Health and Nutrition Examination Survey (NHANES) 1999-2000. *Environ. Health Perspect.* **112** 331–338. https://doi.org/10.1289/ehp.6723

SPEARMAN, C. (1904). General intelligence, objectively determined and measured. *Amer. J. Psychol.* **15** 201–293.

STAFOGGIA, M., BREITNER, S., HAMPEL, R. and BASAGAÑA, X. (2017). Statistical approaches to address multipollutant mixtures and multiple exposures: The state of the science. *Curr. Environ. Health Rep.* **4** 481–490. https://doi.org/10.1007/s40572-017-0162-z

STAHLHUT, R. W., VAN WIJNGAARDEN, E., DYE, T. D., COOK, S. and SWAN, S. H. (2007). Concentrations of urinary phthalate metabolites are associated with increased waist circumference and insulin resistance in adult U.S. males. *Environ. Health Perspect.* **115** 876–882. https://doi.org/10.1289/ehp.9882

STOJANOSKA, M. M., MILOSEVIC, N., MILIC, N. and ABENAVOLI, L. (2017). The influence of phthalates and bisphenol A on the obesity development and glucose metabolism disorders. *Endocrine* **55** 666–681. https://doi.org/10.1007/s12020-016-1158-4

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

SUN, Q., CORNELIS, M. C., TOWNSEND, M. K., TOBIAS, D. K., ELIASSEN, A. H., FRANKE, A. A., HAUSER, R. and HU, F. B. (2014). Association of urinary concentrations of bisphenol A and phthalate metabolites with risk of type 2 diabetes: A prospective investigation in the Nurses' Health Study (NHS) and NHSII cohorts. *Environ. Health Perspect.* **122** 616–623.

TANNER, E. M., BORNEHAG, C.-G. and GENNINGS, C. (2019). Repeated holdout validation for weighted quantile sum regression. *MethodsX* **6** 2855–2860. https://doi.org/10.1016/j.mex.2019.11.008

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

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

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN DER PAS, S., SCOTT, J., CHAKRABORTY, A. and BHATTACHARYA, A. (2016). horseshoe: Implementation of the horseshoe prior. *CRAN Repository*. R package version 0.2.0.

WANG, C., JIANG, B. and ZHU, L. (2021). Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statist. Sinica* **31** 1549–1570. MR4297716 https://doi.org/10.5705/ss.202019.0081

WANG, Y., GHASSABIAN, A., GU, B., AFANASYEVA, Y., LI, Y., TRASANDE, L. and LIU, M. (2023). Semiparametric distributed lag quantile regression for modeling time-dependent exposure mixtures. *Biometrics* **79** 2619–2632. MR4644020 https://doi.org/10.1111/biom.13702

WARREN, J., FUENTES, M., HERRING, A. and LANGLOIS, P. (2012). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics* **68** 1157–1167. MR3040022 https://doi.org/10.1111/j.1541-0420.2012.01774.x

WARREN, J. L., CHANG, H. H., WARREN, L. K., STRICKLAND, M. J., DARROW, L. A. and MULHOLLAND, J. A. (2022). Critical window variable selection for mixtures: Estimating the impact of multiple air pollutants on stillbirth. *Ann. Appl. Stat.* **16** 1633–1652. MR4455894 https://doi.org/10.1214/21-aoas1560

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

WARREN, J. L., STINGONE, J. A., HERRING, A. H. et al. (2016). Bayesian multinomial probit modeling of daily windows of susceptibility for maternal $PM_{2.5}$ exposure and congenital heart defects. *Stat. Med.* **35** 2786–2801. MR3513718 https://doi.org/10.1002/sim.6891

WELTY, L. J., PENG, R. D., ZEGER, S. L. and DOMINICI, F. (2009). Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. *Biometrics* **65** 282–291. MR2665891 https://doi.org/10.1111/j.1541-0420.2007.01039.x

WENZEL, A. G., BROCK, J. W., CRUZE, L., NEWMAN, R. B., UNAL, E. R., WOLF, B. J., SOMERVILLE, S. E. and KUCKLICK, J. R. (2018). Prevalence and predictors of phthalate exposure in pregnant women in Charleston, SC. *Chemosphere* **193** 394–402.

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.*.

WHYATT, R. M., LIU, X., RAUH, V. A., CALAFAT, A. M., JUST, A. C., HOEPNER, L., DIAZ, D., QUINN, J., ADIBI, J. et al. (2012). Maternal prenatal urinary phthalate metabolite concentrations and child mental, psychomotor, and behavioral development at 3 years of age. *Environ. Health Perspect.* **120** 290–295.

WILSON, A., HSU, H.-H. L., CHIU, Y.-H. M., WRIGHT, R. O., WRIGHT, R. J. and COULL, B. A. (2022). Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies. *Ann. Appl. Stat.* **16** 1090–1110. MR4438825 https://doi.org/10.1214/21-aoas1533

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

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

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 798–820
https://doi.org/10.1214/24-AOAS1989

FUNCTIONAL CLUSTERING FOR LONGITUDINAL ASSOCIATIONS BETWEEN SOCIAL DETERMINANTS OF HEALTH AND STROKE MORTALITY IN THE U.S.

BY FANGZHI LUO[1,a], JIANBIN TAN[2,c], DONGLAN ZHANG[3,d], HUI HUANG[4,e] AND YE SHEN[1,b]

[1]*Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia,* [a]*fl81224@uga.edu,* [b]*yeshen@uga.edu*

[2]*Department of Biostatistics and Bioinformatics, School of Medicine, Duke University,* [c]*jianbin.tan@duke.edu*

[3]*Department of Foundations of Medicine, NYU Grossman Long Island School of Medicine,* [d]*Donglan.Zhang@nyulangone.org*

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

Understanding the longitudinally changing associations between Social Determinants of Health (SDOH) and stroke mortality is essential for effective stroke management. Previous studies have uncovered significant regional disparities in the associations between SDOH and stroke mortality. However, existing studies have not utilized longitudinal associations to develop data-driven methods for regional division in stroke control. To fill this gap, we propose a novel clustering method to analyze SDOH-stroke mortality associations across U.S. counties. To enhance the interpretability of the clustering outcomes, we introduce a novel regularized expectation-maximization algorithm equipped with sparsity- and smoothness-pursued penalties, aiming at simultaneous clustering and variable selection in longitudinal associations. As a result, we can identify key SDOH that contributes to longitudinal changes in stroke mortality. This facilitates the clustering of U.S. counties based on the associations between these SDOH and stroke mortality. The effectiveness of our proposed method is demonstrated through extensive numerical studies. By applying our method to longitudinal data on SDOH and stroke mortality at the county level, we identify 18 important SDOH for stroke mortality and divide the U.S. counties into two clusters based on these selected SDOH. Our findings unveil complex regional heterogeneity in the longitudinal associations between SDOH and stroke mortality, providing valuable insights into region-specific SDOH adjustments for mitigating stroke mortality.

REFERENCES

ANEIROS, G., NOVO, S. and VIEU, P. (2022). Variable selection in functional regression models: A review. *J. Multivariate Anal.* **188** Paper No. 104871. MR4353868 https://doi.org/10.1016/j.jmva.2021.104871

BENJAMIN, E. J., MUNTNER, P., ALONSO, A., BITTENCOURT, M. S., CALLAWAY, C. W., CARSON, A. P., CHAMBERLAIN, A. M., CHANG, A. R., CHENG, S. et al. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. *Circulation* **139** e56–e528.

BREHENY, P. and HUANG, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **25** 173–187. MR3306699 https://doi.org/10.1007/s11222-013-9424-2

CAI, T. T., MA, J. and ZHANG, L. (2019). CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *Ann. Statist.* **47** 1234–1267. MR3911111 https://doi.org/10.1214/18-AOS1711

CAI, X., XUE, L. and CAO, J. (2022). Variable selection for multiple function-on-function linear regression. *Statist. Sinica* **32** 1435–1465. MR4449901 https://doi.org/10.5705/ss.202020.0473

CHIOU, J.-M. and LI, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 679–699. MR2370075 https://doi.org/10.1111/j.1467-9868.2007.00605.x

Key words and phrases. Model-based clustering, functional regression model, variable selection, regularized expectation-maximization algorithm, stroke mortality.

DING, J., TAROKH, V. and YANG, Y. (2018). Model selection techniques: An overview. *IEEE Signal Process. Mag.* **35** 16–34.

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

FRIEDEN, T. R. et al. (2013). CDC health disparities and inequalities report-United States, 2013. Foreword. *MMWR Supplements* **62** 1–2.

GEBREAB, S. Y., DAVIS, S. K., SYMANZIK, J., MENSAH, G. A., GIBBONS, G. H. and DIEZ-ROUX, A. V. (2015). Geographic variations in cardiovascular health in the United States: Contributions of state- and individual-level factors. *J. Amer. Heart Assoc.* **4** e001673. https://doi.org/10.1161/JAHA.114.001673

GHOSAL, R., MAITY, A., CLARK, T. and LONGO, S. B. (2020). Variable selection in functional linear concurrent regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 565–587. MR4098962 https://doi.org/10.1111/rssc.12408

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

GRÖMPING, U. (2007). Relative importance for linear regression in R: The package relaimpo. *J. Stat. Softw.* **17** 1–27.

HAVRANEK, E. P., MUJAHID, M. S., BARR, D. A., BLAIR, I. V., COHEN, M. S., CRUZ-FLORES, S., DAVEY-SMITH, G., DENNISON-HIMMELFARB, C. R., LAUER, M. S. et al. (2015). Social determinants of risk and outcomes for cardiovascular disease: A scientific statement from the American Heart Association. *Circulation* **132** 873–898.

HE, J., ZHU, Z., BUNDY, J. D., DORANS, K. S., CHEN, J. and HAMM, L. L. (2021). Trends in cardiovascular risk factors in US adults by race and ethnicity and socioeconomic status, 1999-2018. *JAMA* **326** 1286–1298.

HEYDEN, S., TYROLER, H. A., HEISS, G., HAMES, C. G. and BARTEL, A. (1978). Coffee consumption and mortality. Total mortality, stroke mortality, and coronary heart disease mortality. *Arch. Intern. Med.* **138** 1472–1475.

HOLLOWAY, R. G., ARNOLD, R. M., CREUTZFELDT, C. J., LEWIS, E. F., LUTZ, B. J., MCCANN, R. M., RABINSTEIN, A. A., SAPOSNIK, G., SHETH, K. N. et al. (2014). Palliative and end-of-life care in stroke: A statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **45** 1887–1916.

HOWARD, G. and HOWARD, V. J. (2020). Twenty years of progress toward understanding the stroke belt. *Stroke* **51** 742–750. https://doi.org/10.1161/STROKEAHA.119.024155

HUANG, J. Z., SHEN, H. and BUJA, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *J. Amer. Statist. Assoc.* **104** 1609–1620. MR2750581 https://doi.org/10.1198/jasa.2009.tm08024

HUANG, J. Z., WU, C. O. and ZHOU, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349 https://doi.org/10.1093/biomet/89.1.111

IMOISILI, O. E. (2024). Prevalence of stroke—behavioral risk factor surveillance system, United States, 2011–2022. *Morb. Mort. Wkly. Rep.* **73**.

JACQUES, J. and PREDA, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing* **112** 164–171.

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

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

KAPRAL, M. K., AUSTIN, P. C., JEYAKUMAR, G., HALL, R., CHU, A., KHAN, A. M., JIN, A. Y., MARTIN, C., MANUEL, D. et al. (2019). Rural-urban differences in stroke risk factors, incidence, and mortality in people with and without prior stroke: The CANHEART stroke study. *Circ. Cardiovasc. Qual. Outcomes* **12** e004973.

KARP, D. N., WOLFF, C. S., WIEBE, D. J., BRANAS, C. C., CARR, B. G. and MULLEN, M. T. (2016). Reassessing the stroke belt: Using small area spatial statistics to identify clusters of high stroke mortality in the United States. *Stroke* **47** 1939–1942. https://doi.org/10.1161/STROKEAHA.116.012997

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

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

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

KOMÁREK, A. and KOMÁRKOVÁ, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *Ann. Appl. Stat.* **7** 177–200. MR3086415 https://doi.org/10.1214/12-AOAS580

KONG, D., XUE, K., YAO, F. and ZHANG, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103** 147–159. MR3465827 https://doi.org/10.1093/biomet/asv062

KOTON, S., SCHNEIDER, A. L. C., ROSAMOND, W. D., SHAHAR, E., SANG, Y., GOTTESMAN, R. F. and CORESH, J. (2014). Stroke incidence and mortality trends in US communities, 1987 to 2011. *JAMA* **312** 259–268. https://doi.org/10.1001/jama.2014.7692

KOWARIK, A. and TEMPL, M. (2016). Imputation with the R package VIM. *J. Stat. Softw.* **74** 1–16.

LABARTHE, D., GROVER, B., GALLOWAY, J., GORDON, L., MOFFATT, S., PEARSON, T., SCHOEBERL, M. and SIDNEY, S. (2014). *The Public Health Action Plan to Prevent Heart Disease and Stroke*: *Ten-Year Update*. National Forum for Heart Disease and Stroke Prevention, Washington, DC.

LANSKA, D. J. and KULLER, L. H. (1995). The geography of stroke mortality in the United States and the concept of a stroke belt. *Stroke* **26** 1145–1149. https://doi.org/10.1161/01.str.26.7.1145

LI, F. and SANG, H. (2019). Spatial homogeneity pursuit of regression coefficients for large datasets. *J. Amer. Statist. Assoc.* **114** 1050–1062. MR4011757 https://doi.org/10.1080/01621459.2018.1529595

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

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

LU, Z. and SONG, X. (2012). Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. *Stat. Med.* **31** 544–560. MR2892338 https://doi.org/10.1002/sim.4420

LUO, F., TAN, J., ZHANG, D., HUANG, H. and SHEN, Y. (2025). Supplement to "Functional clustering for longitudinal associations between social determinants of health and stroke mortality in the U.S.." https://doi.org/10.1214/24-AOAS1989SUPPA, https://doi.org/10.1214/24-AOAS1989SUPPB

MAWHORTER, S., CRIMMINS, E. M. and AILSHIRE, J. A. (2023). Housing and cardiometabolic risk among older renters and homeowners. *Hous Stud* **38** 1342–1364. https://doi.org/10.1080/02673037.2021.1941792

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

MCNICHOLAS, P. D. and MURPHY, T. B. (2010). Model-based clustering of longitudinal data. *Canad. J. Statist.* **38** 153–168. MR2676935 https://doi.org/10.1002/cjs.10047

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443 https://doi.org/10.1214/09-AOS692

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

MOZAFFARIAN, D., BENJAMIN, E. J., GO, A. S., ARNETT, D. K., BLAHA, M. J., CUSHMAN, M., DE FERRANTI, S., DESPRÉS, J.-P., FULLERTON, H. J. et al. (2015). Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation* **131** e29–e322.

PENG, J. and MÜLLER, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Ann. Appl. Stat.* **2** 1056–1077. MR2516804 https://doi.org/10.1214/08-AOAS172

PERPEROGLOU, A., SAUERBREI, W., ABRAHAMOWICZ, M. and SCHMID, M. (2019). A review of spline function procedures in R. *BMC Med. Res. Methodol.* **19** 1–16.

POWELL-WILEY, T. M., BAUMER, Y., BAAH, F. O., BAEZ, A. S., FARMER, N., MAHLOBO, C. T., PITA, M. A., POTHARAJU, K. A., TAMURA, K. et al. (2022). Social determinants of cardiovascular disease. *Circ. Res.* **130** 782–799. https://doi.org/10.1161/CIRCRESAHA.121.319811

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. *Springer Series in Statistics*. Springer, New York. MR1910407 https://doi.org/10.1007/b98886

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

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RECORD, N. B., ONION, D. K., PRIOR, R. E., DIXON, D. C., RECORD, S. S., FOWLER, F. L., CAYER, G. R., AMOS, C. I. and PEARSON, T. A. (2015). Community-wide cardiovascular disease prevention programs and health outcomes in a rural county, 1970-2010. *JAMA* **313** 147–155.

RESHETNYAK, E., NTAMATUNGIRO, M., PINHEIRO, L. C., HOWARD, V. J., CARSON, A. P., MARTIN, K. D. and SAFFORD, M. M. (2020). Impact of multiple social determinants of health on incident stroke. *Stroke* **51** 2445–2453. https://doi.org/10.1161/STROKEAHA.120.028530

RODGERS, J., BRIESACHER, B. A., WALLACE, R. B., KAWACHI, I., BAUM, C. F. and KIM, D. (2019). County-level housing affordability in relation to risk factors for cardiovascular disease among middle-aged adults: The national longitudinal survey of youths 1979. *Health Place* **59** 102194.

SARTÓRIO, V. S. and FONSECA, T. C. (2020). Dynamic clustering of time series data. ArXiv preprint. Available at arXiv:2002.01890.

SHAH, B. R., KHAN, N. A., O'DONNELL, M. J. and KAPRAL, M. K. (2015). Impact of language barriers on stroke care and outcomes. *Stroke* **46** 813–818. https://doi.org/10.1161/STROKEAHA.114.007929

SKOLARUS, L. E., SHARRIEF, A., GARDENER, H., JENKINS, C. and BODEN-ALBALA, B. (2020). Considerations in addressing social determinants of health to reduce racial/ethnic disparities in stroke outcomes in the United States. *Stroke* **51** 3433–3439. https://doi.org/10.1161/STROKEAHA.120.030426

SON, H., ZHANG, D., SHEN, Y., JAYSING, A., ZHANG, J., CHEN, Z., MU, L., LIU, J., RAJBHANDARI-THAPA, J. et al. (2023). Social determinants of cardiovascular health: A longitudinal analysis of cardiovascular disease mortality in US counties from 2009 to 2018. *J. Amer. Heart Assoc.* **12** e026940.

TANG, X. and QU, A. (2016). Mixture modeling for longitudinal data. *J. Comput. Graph. Statist.* **25** 1117–1137. MR3572031 https://doi.org/10.1080/10618600.2015.1092979

TSAO, C. W., ADAY, A. W., ALMARZOOQ, Z. I., ALONSO, A., BEATON, A. Z., BITTENCOURT, M. S., BOEHME, A. K., BUXTON, A. E., CARSON, A. P. et al. (2022). Heart disease and stroke statistics-2022 update: A report from the American Heart Association. *Circulation* **145** e153–e639. https://doi.org/10.1161/CIR.0000000000001052

TSAO, C. W., ADAY, A. W., ALMARZOOQ, Z. I., ANDERSON, C. A. M., ARORA, P., AVERY, C. L., BAKER-SMITH, C. M., BEATON, A. Z., BOEHME, A. K. et al. (2023). Heart disease and stroke statistics-2023 update: A report from the American heart association. *Circulation* **147** e93–e621. https://doi.org/10.1161/CIR.0000000000001123

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

VILLABLANCA, A. C., SLEE, C., LIANOV, L. and TANCREDI, D. (2016). Outcomes of a clinic-based educational intervention for cardiovascular disease prevention by race, ethnicity, and urban/rural status. *J. Women's Health* **25** 1174–1186.

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

WANG, L., CHEN, G. and LI, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486–1494.

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204 https://doi.org/10.1198/016214508000000788

WANG, X., LIU, L. Y.-F. and ZHU, H. (2023). Functional finite mixture regression models. *Statist. Sinica* **33** 2087–2115. MR4607201

YAO, F., FU, Y. and LEE, T. C. (2011). Functional mixture regression. *Biostatistics* **12** 341–353.

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

YI, X. and CARAMANIS, C. (2015). Regularized em algorithms: A unified framework and statistical guarantees. *Adv. Neural Inf. Process. Syst.* **28**.

ZELKO, A., SALERNO, P. R. V. O., AL-KINDI, S., HO, F., ROCHA, F. P., NASIR, K., RAJAGOPALAN, S., DEO, S. and SATTAR, N. (2023). Geographically weighted modeling to explore social and environmental factors affecting county-level cardiovascular mortality in people with diabetes in the United States: A cross-sectional analysis. *Am. J. Cardiol.* **209** 193–198. https://doi.org/10.1016/j.amjcard.2023.09.084

ZENG, L. and XIE, J. (2014). Group variable selection via SCAD-L_2. *Statistics* **48** 49–66. MR3175757 https://doi.org/10.1080/02331888.2012.719513

ZHOU, H., YAO, F. and ZHANG, H. (2023). Functional linear regression for discretely observed data: From ideal to reality. *Biometrika* **110** 381–393. MR4588359 https://doi.org/10.1093/biomet/asac053

ZHU, H., LI, R. and KONG, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40** 2634–2666. MR3097615 https://doi.org/10.1214/12-AOS1045

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

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470 https://doi.org/10.1214/08-AOS625

The Annals of Applied Statistics
2025, Vol. 19, No. 1, 821–845
https://doi.org/10.1214/24-AOAS1990

NONPARAMETRIC CAUSAL DECOMPOSITION OF GROUP DISPARITIES

BY ANG YU[a] AND FELIX ELWERT[b]

Department of Sociology, University of Wisconsin-Madison, [a]*ayu33@wisc.edu,* [b]*elwert@wisc.edu*

We introduce a new nonparametric causal decomposition approach that identifies the mechanisms by which a treatment variable contributes to a group-based outcome disparity. Our approach distinguishes three mechanisms: group differences in: (1) treatment prevalence, (2) average treatment effects, and (3) selection into treatment based on individual-level treatment effects. Our approach reformulates classic Kitagawa–Blinder–Oaxaca decompositions in causal and nonparametric terms, complements causal mediation analysis by explaining group disparities instead of group effects, and isolates conceptually distinct mechanisms conflated in recent random equalization decompositions. In contrast to all prior approaches, our framework uniquely identifies differential selection into treatment as a novel disparity-generating mechanism. Our approach can be used for both the retrospective causal explanation of disparities and the prospective planning of interventions to change disparities. We present both an unconditional and a conditional decomposition, where the latter quantifies the contributions of the treatment within levels of certain covariates. We develop nonparametric estimators that are $\sqrt{n}$-consistent, asymptotically normal, semiparametrically efficient, and multiply robust. We apply our approach to analyze the mechanisms by which college graduation causally contributes to intergenerational income persistence (the disparity in adult income between the children of high- vs. low-income parents). Empirically, we demonstrate a previously undiscovered role played by the new selection component in intergenerational income persistence.

REFERENCES

AVIN, C., SHPITSER, I. and PEARL, J. (2005). Identifiability of path-specific effects. In *Proceedings of International Joint Conference on Artificial Intelligence* 357–363.

BAILEY, M. and DYNARSKI, S. (2011). Gains and gaps: Changing inequality in U.S. college entry and completion. Technical Report No. w17633, National Bureau of Economic Research, Cambridge, MA.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

BLAU, P. M. and DUNCAN, O. D. (1967). *The American Occupational Structure*. Free Press, New York.

BLINDER, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *J. Hum. Resour.* **8** 436.

BLOOME, D., DYER, S. and ZHOU, X. (2018). Educational inequality, educational expansion, and intergenerational income persistence in the United States. *Am. Sociol. Rev.* **83** 1215–1253.

BRADY, D., FINNIGAN, R. M. and HÜBGEN, S. (2017). Rethinking the risks of poverty: A framework for analyzing prevalences and penalties. *Am. J. Sociol.* **123** 740–786.

BRAND, J. E. and XIE, Y. (2010). Who benefits most from college?: Evidence for negative selection in heterogeneous economic returns to higher education. *Am. Sociol. Rev.* **75** 273–302.

BRAND, J. E., XU, J., KOCH, B. and GERALDO, P. (2021). Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociol. Method.* **51** 189–223.

BREEN, R., ed. (2004). *Social Mobility in Europe*. Oxford Univ. Press, London.

CHA, Y., WEEDEN, K. A. and SCHNABEL, L. (2023). Is the gender wage gap really a family wage gap in disguise? *Am. Sociol. Rev.* **88** 972–1001.

CHENG, S., BRAND, J. E., ZHOU, X., XIE, Y. and HOUT, M. (2021). Heterogeneous returns to college over the life course. *Sci. Adv.* **7** eabg7641.

Key words and phrases. Social inequality, mediation analysis, double machine learning, income mobility.

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

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and LUO, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica* **86** 1911–1938. MR3901242 https://doi.org/10.3982/ECTA14415

DÍAZ, I., HEJAZI, N. S., RUDOLPH, K. E. and VAN DER LAAN, M. J. (2021). Nonparametric efficient causal mediation with intermediate confounders. *Biometrika* **108** 627–641. MR4298768 https://doi.org/10.1093/biomet/asaa085

DIDELEZ, V., DAWID, A. P. and GENELETTI, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. S. Richardson, eds.) 138–146. AUAI Press, Arlington, VA.

DIDERICHSEN, F., HALLQVIST, J. and WHITEHEAD, M. (2019). Differential vulnerability and susceptibility: How to make use of recent development in our understanding of mediation and interaction to tackle health inequalities. *Int. J. Epidemiol.* **48** 268–274. https://doi.org/10.1093/ije/dyy167

FEATHERMAN, D. L. and HAUSER, R. M. (1978). *Opportunity and Change*. *Studies in Population*. Academic Press, San Diego.

FIEL, J. E. (2020). Great equalizer or great selector? Reconsidering education as a moderator of intergenerational transmissions. *Soc. Educ.* **93** 353–371.

FORTIN, N., LEMIEUX, T. and FIRPO, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics* **4** 1–102. Elsevier, Amsterdam.

GENELETTI, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 199–215. MR2325272 https://doi.org/10.1111/j.1467-9868.2007.00584.x

HECKMAN, J. J., HUMPHRIES, J. E. and VERAMENDI, G. (2018). Returns to education: The causal effects of education on earnings, health, and smoking. *J. Polit. Econ.* **126** 50.

HECKMAN, J. J., STIXRUD, J. and URZUA, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econ.* **24** 411–482.

HECKMAN, J. J. and VYTLACIL, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* **73** 669–738. MR2135141 https://doi.org/10.1111/j.1468-0262.2005.00594.x

HINES, O., DUKES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2022). Demystifying statistical learning based on efficient influence functions. *Amer. Statist.* **76** 292–304. MR4453533 https://doi.org/10.1080/00031305.2021.2021984

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

HOUT, M. (2012). Social and economic returns to college education in the United States. *Annu. Rev. Sociol.* **38** 379–400.

HOWE, C. J., NAPRAVNIK, S., COLE, S. R., KAUFMAN, J. S., ADIMORA, A. A., ELSTON, B., ERON, J. J. and MUGAVERO, M. J. (2014). African American race and HIV virological suppression: Beyond disparities in clinic attendance. *Amer. J. Epidemiol.* **179** 1484–1492. https://doi.org/10.1093/aje/kwu069

HUBER, M. (2015). Causal pitfalls in the decomposition of wage gaps. *J. Bus. Econom. Statist.* **33** 179–191. MR3337056 https://doi.org/10.1080/07350015.2014.937437

ISHIDA, H., MULLER, W. and RIDGE, J. M. (1995). Class origin, class destination, and education: A cross-national study of ten industrial nations. *Am. J. Sociol.* **101** 145–193.

JACKSON, J. W. (2021). Meaningful causal decompositions in health equity research: Definition, identification, and estimation through a weighting framework. *Epidemiology* **32** 282–290. https://doi.org/10.1097/EDE.0000000000001319

JACKSON, J. W. and VANDERWEELE, T. J. (2018). Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* **29** 825–835. https://doi.org/10.1097/EDE.0000000000000901

KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. Available at arXiv:2203.06469 [stat].

KITAGAWA, E. M. (1955). Components of a difference between two rates. *J. Amer. Statist. Assoc.* **50** 1168–1194.

KLINE, P. (2011). Oaxaca-blinder as a reweighting estimator. *Amer. Econ. Rev.* **101** 532–537.

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multilevel data. *Stat. Med.* **32** 3373–3387. MR3074363 https://doi.org/10.1002/sim.5786

LUNDBERG, I. (2024). The gap-closing estimand: A causal approach to study interventions that close disparities across social categories. *Sociol. Methods Res.* **53** 507–570. MR4736949 https://doi.org/10.1177/00491241211055769

LUNDBERG, I., JOHNSON, R. and STEWART, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.* **86** 532–565.

MALINSKY, D. (2018). Intervening on structure. *Synthese* **195** 2295–2312. MR3784956 https://doi.org/10.1007/s11229-017-1341-z

MARE, R. D. (1980). Social background and school continuation decisions. *J. Amer. Statist. Assoc.* **75** 295–305.

NAIMI, A. I., SCHNITZER, M. E., MOODIE, E. E. M. and BODNAR, L. M. (2016). Mediation analysis for health disparities research. *Amer. J. Epidemiol.* **184** 315–324. https://doi.org/10.1093/aje/kwv329

OAXACA, R. (1973). Male-female wage differentials in urban labor markets. *Internat. Econom. Rev.* **14** 693.

OPACIC, A., WEI, L. and ZHOU, X. (2023). Disparity analysis: A tale of two approaches. Unpublished manuscript.

PARK, C. and KANG, H. (2024). A groupwise approach for inferring heterogeneous treatment effects in causal inference. *J. Roy. Statist. Soc. Ser. A* **187** 374–392. MR4745596 https://doi.org/10.1093/jrsssa/qnad125

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intel Ligence* 411–20. Morgan Kaufmann, San Francisco, CA.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

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

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

RUBIN, D. B. (1986). Comment: The design and analysis of gold standard randomized experiments [MR2655714]. *J. Amer. Statist. Assoc.* **81** 961–962.

SEMENOVA, V. and CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* **24** 264–289. MR4281225 https://doi.org/10.1093/ectj/utaa027

SUDHARSANAN, N. and BIJLSMA, M. J. (2021). Educational note: Causal decomposition of population health differences using Monte Carlo integration and the g-formula. *Int. J. Epidemiol.* **50** 2098–2107. https://doi.org/10.1093/ije/dyab090

VAN DER LAAN, M. J. (2006). Statistical inference for variable importance. *Int. J. Biostat.* **2** Art. 2. MR2275897 https://doi.org/10.2202/1557-4679.1008

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford Univ. Press, London.

VANDERWEELE, T. J. (2014). A unification of mediation and interaction: A 4-way decomposition. *Epidemiology* **25** 749–761. https://doi.org/10.1097/EDE.0000000000000121

VANDERWEELE, T. J. and ROBINSON, W. R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25** 473–484. https://doi.org/10.1097/EDE.0000000000000105

VANDERWEELE, T. J. and TCHETGEN, E. J. T. (2014). Attributing effects to interactions. *Epidemiology* **25** 711–722. https://doi.org/10.1097/EDE.0000000000000096

WARD, J. B., GARTNER, D. R., KEYES, K. M., FLISS, M. D., MCCLURE, E. S. and ROBINSON, W. R. (2019). How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies. *Ann. Epidemiol.* **29** 1–7. https://doi.org/10.1016/j.annepidem.2018.09.007

YAMAGUCHI, K. (2015). Decomposition of gender or racial inequality with endogenous intervening covariates: An extension of the DiNardo–Fortin–Lemieux method. *Sociol. Method.* **45** 388–428.

YU, A. (2024). cdgd: Causal Decomposition of Group Disparities. R package version 0.3.5.

YU, A. and ELWERT, F. (2025). Supplement to "Nonparametric Causal Decomposition of Group Disparities." https://doi.org/10.1214/24-AOAS1990SUPPA, https://doi.org/10.1214/24-AOAS1990SUPPB

YU, A., GE, L. and ELWERT, F. (2024). When do natural mediation effects differ from their randomized interventional analogues: Test and theory. Available at arXiv:2407.02671 [stat].

YU, A., PARK, C., KANG, H. and FLETCHER, J. (2021). Leveraging machine learning to estimate effect modification. Preprint, SocArXiv.

ZHOU, X. (2019). Equalization or selection? Reassessing the "meritocratic power" of a college degree in intergenerational income mobility. *Am. Sociol. Rev.* **84** 459–485.

ZHOU, X. (2024). Attendance, completion, and heterogeneous returns to college: A causal mediation approach. *Sociol. Methods Res.* **53** 1136–1166. MR4783051 https://doi.org/10.1177/00491241221113876

ZIOL-GUEST, K. M. and LEE, K. T. H. (2016). Parent income–based gaps in schooling: Cross-cohort trends in the NLSYs and the PSID. *AERA Open* **2** 233285841664583.