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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

THE ANNALS OF APPLIED STATISTICS Vol. 17, No. 4, pp. 2701–3588 December 2023

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

IMS PUBLICATIONS

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 17, Number 4, December 2023. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to *The Annals of Applied Statistics*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2701–2722
https://doi.org/10.1214/22-AOAS1700

A MULTIAGENT REINFORCEMENT LEARNING FRAMEWORK FOR OFF-POLICY EVALUATION IN TWO-SIDED MARKETS

BY CHENGCHUN SHI[1,a], RUNZHE WAN[2,b], GE SONG[3,d], SHIKAI LUO[3,e], HONGTU ZHU[4,f] AND RUI SONG[2,c]

[1]*London School of Economics and Political Science,* [a]*c.shi7@lse.ac.uk*
[2]*North Carolina State University,* [b]*rwan@ncsu.edu,* [c]*rsong@ncsu.edu*
[3]*Didi Chuxing,* [d]*songge@didiglobal.com,* [e]*luoshikai@didiglobal.com*
[4]*The University of North Carolina at Chapel Hill,* [f]*htzhu@email.unc.edu*

The two-sided markets, such as ride-sharing companies, often involve a group of subjects who are making sequential decisions across time and/or location. With the rapid development of smart phones and internet of things, they have substantially transformed the transportation landscape of human beings. In this paper we consider large-scale fleet management in ride-sharing companies that involve multiple units in different areas receiving sequences of products (or treatments) over time. Major technical challenges, such as policy evaluation, arise in those studies because: (i) spatial and temporal proximities induce interference between locations and times, and (ii) the large number of locations results in the curse of dimensionality. To address both challenges simultaneously, we introduce a multiagent reinforcement learning (MARL) framework for carrying policy evaluation in these studies. We propose novel estimators for mean outcomes under different products that are consistent despite the high dimensionality of state-action space. The proposed estimator works favorably in simulation experiments. We further illustrate our method using a real dataset obtained from a two-sided marketplace company to evaluate the effects of applying different subsidizing policies. A Python implementation of our proposed method is available in the Supplementary Material and also at https://github.com/RunzheStat/CausalMARL.

REFERENCES

ARMSTRONG, M. (2006). Competition in two-sided markets. *Rand J. Econ*. **37** 668–691.

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

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771 https://doi.org/10.3982/ECTA12723

BHATTACHARYA, R., MALINSKY, D. and SHPITSER, I. (2019). Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*: *Proceedings of the... Conference*. *Conference on Uncertainty in Artificial Intelligence* **2019**. NIH Public Access.

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

BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc*. **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

CAI, H., SHI, C., SONG, R. and LU, W. (2021). Deep jump learning for off-policy evaluation in continuous treatment settings. *Adv. Neural Inf. Process. Syst*. **34** 15285–15300.

CHAKRABORTY, B., LABER, E. B. and ZHAO, Y.-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials* **11** 408–417.

Key words and phrases. Multiagent system, reinforcement learning, spatiotemporal studies, policy evaluation.

CHAKRABORTY, B., MURPHY, S. and STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* **19** 317–343. MR2757118 https://doi.org/10.1177/0962280209105013

CHEN, E. Y., HU, Z. T., SONG, R. and JORDAN, M. I. (2020). Heterogeneous reinforcement learning with offline data: Estimation and inference.

CHEN, X. and QI, Z. (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. ArXiv preprint. Available at arXiv:2201.06169.

DEMPSEY, W., LIAO, P., KUMAR, S. and MURPHY, S. A. (2020). The stratified micro-randomized trial design: Sample size considerations for testing nested causal effects of time-varying treatments. *Ann. Appl. Stat.* **14** 661–684. MR4117824 https://doi.org/10.1214/19-AOAS1293

DUDÍK, M., ERHAN, D., LANGFORD, J. and LI, L. (2014). Doubly robust policy evaluation and optimization. *Statist. Sci.* **29** 485–511. MR3300356 https://doi.org/10.1214/14-STS500

ERTEFAIE, A. (2014). Constructing dynamic treatment regimes in infinite-horizon settings. ArXiv preprint. Available at arXiv:1406.0764.

FANG, E. X. WANG, Z. and WANG, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *J. Amer. Statist. Assoc.* To appear.

FARAHMAND, A., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *J. Mach. Learn. Res.* **17** Paper No. 139. MR3555030

FRENKEN, K. and SCHOR, J. (2017). Putting the sharing economy into perspective. *Environmental Innovation and Societal Transitions* **23** 3–10.

FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2007). Kernel measures of conditional dependence. In *NIPS* **20** 489–496.

HAGIU, A. and WRIGHT, J. (2019). The status of workers and platforms in the sharing economy. *J. Econ. Manag. Strategy* **28** 97–108.

HALLORAN, M. E. and HUDGENS, M. G. (2016). Dependent happenings: A recent methodological review. *Curr. Epidemiol. Rep.* **3** 297–305.

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

HU, X., QIAN, M., CHENG, B. and CHEUNG, Y. K. (2021). Personalized policy learning using longitudinal mobile health data. *J. Amer. Statist. Assoc.* **116** 410–420. MR4227703 https://doi.org/10.1080/01621459.2020.1785476

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

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

JIANG, N. and LI, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* 652–661.

JIN, S. T., KONG, H., WU, R. and SUI, D. Z. (2018). Ridesourcing, the sharing economy, and the future of cities. *Cities* **76** 96–104.

KALLUS, N. and UEHARA, M. (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.* **70** 3282–3302. MR4538517

LABER, E. B., MEYER, N. J., REICH, B. J., PACIFICI, K., COLLAZO, J. A. and DRAKE, J. M. (2018). Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 743–789. MR3832250 https://doi.org/10.1111/rssc.12266

LI, B., ZHANG, D., SUN, L., CHEN, C., LI, S., QI, G. and YANG, Q. (2011). Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. In 2011 *IEEE International Conference on Pervasive Computing and Communications Workshops* (*PERCOM Workshops*) 63–68. IEEE Press, New York.

LI, M., SHI, C., WU, Z. and FRYZLEWICZ, P. (2022a). Reinforcement learning in possibly nonstationary environments. ArXiv preprint. Available at arXiv:2203.01707.

LI, Y., WANG, C.-H., CHENG, G. and SUN, W. W. (2022b). Rate-optimal contextual online matching bandit. ArXiv preprint. Available at arXiv:2205.03699.

LIAO, P., KLASNJA, P. and MURPHY, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *J. Amer. Statist. Assoc.* **116** 382–391. MR4227701 https://doi.org/10.1080/01621459.2020.1807993

LIAO, P., QI, Z., WAN, R., KLASNJA, P. and MURPHY, S. A. (2022). Batch policy learning in average reward Markov decision processes. *Ann. Statist.* **50** 3364–3387. MR4524500 https://doi.org/10.1214/22-aos2231

LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems* 5356–5366.

LLOYD, E. H. (1977). Reservoirs with seasonally varying Markovian inflows and their first passage times.
LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919
LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384
MATSOUAKA, R. A., LI, J. and CAI, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics* **70** 489–499. MR3261768 https://doi.org/10.1111/biom.12179
MENG, H., ZHAO, Y.-Q., FU, H. and QIAO, X. (2020). Near-optimal individualized treatment recommendations. *J. Mach. Learn. Res.* **21** Paper No. 183. MR4209469
MIAO, F., HAN, S., LIN, S., STANKOVIC, J. A., ZHANG, D., MUNIR, S., HUANG, H., HE, T. and PAPPAS, G. J. (2016). Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach. *IEEE Trans. Autom. Sci. Eng.* **13** 463–478.
MO, W., QI, Z. and LIU, Y. (2021). Learning optimal distributionally robust individualized treatment rules. *J. Amer. Statist. Assoc.* **116** 659–674. MR4270012 https://doi.org/10.1080/01621459.2020.1796359
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
NACHUM, O., CHOW, Y., DAI, B. and LI, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. ArXiv preprint. Available at arXiv:1906.04733.
NING, B., GHOSAL, S. and THOMAS, J. (2019). Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Anal.* **14** 1–28. MR3910036 https://doi.org/10.1214/18-BA1102
PUTERMAN, M. L. (1994). *Markov Decision Processes*: *Discrete Stochastic Dynamic Programming*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1270015
QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864
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
ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11
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: Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.
RYSMAN, M. (2009). The economics of two-sided markets. *J. Econ. Perspect.* **23** 125–143.
SHI, C., FAN, A., SONG, R. and LU, W. (2018). High-dimensional *A*-learning for optimal dynamic treatment regimes. *Ann. Statist.* **46** 925–957. MR3797992 https://doi.org/10.1214/17-AOS1570
SHI, C., LU, W. and SONG, R. (2020). Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *J. Mach. Learn. Res.* **21** Paper No. 176. MR4209462
SHI, C., SONG, R., LU, W. and FU, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 681–702. MR3849339 https://doi.org/10.1111/rssb.12273
SHI, C., WAN, R., CHERNOZHUKOV, V. and SONG, R. (2021). Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning* 9580–9591. PMLR.
SHI, C., WAN, R., SONG, G., LUO, S., ZHU, H. and SONG, R. (2023). Supplement to "A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets." https://doi.org/10.1214/22-AOAS1700SUPP
SHI, C., WAN, R., SONG, R., LU, W. and LENG, L. (2020). Does the Markov decision process fit the data: Testing for the Markov property in sequential decision making. In *International Conference on Machine Learning* 8807–8817. PMLR.
SHI, C., WANG, X., LUO, S., ZHU, H., YE, J. and SONG, R. (2022a). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *J. Amer. Statist. Assoc.* 1–13.
SHI, C., ZHANG, S., LU, W. and SONG, R. (2022b). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 765–793. MR4460575
SONG, R., WANG, W., ZENG, D. and KOSOROK, M. R. (2015). Penalized *Q*-learning for dynamic treatment regimens. *Statist. Sinica* **25** 901–920. MR3409730
SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

TANG, Z., FENG, Y., LI, L., ZHOU, D. and LIU, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. ArXiv preprint. Available at arXiv:1910.07186.
TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779
THOMAS, P. and BRUNSKILL, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning* 2139–2148. PMLR.
THOMAS, P. S., THEOCHAROUS, G. and GHAVAMZADEH, M. (2015). High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. In *International Conference on Machine Learning* 1489–1497.
UEHARA, M., HUANG, J. and JIANG, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning* 9659–9668. PMLR.
WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839
WANG, J., QI, Z. and WONG, R. K. (2021). Projected state-action balancing weights for offline reinforcement learning. ArXiv preprint. Available at arXiv:2109.04640.
WANG, L., YANG, Z. and WANG, Z. (2020). Provably efficient causal reinforcement learning with confounded observational data. ArXiv preprint. Available at arXiv:2006.12311.
WANG, L., ZHOU, Y., SONG, R. and SHERWOOD, B. (2018). Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* **113** 1243–1254. MR3862354 https://doi.org/10.1080/01621459.2017.1330204
WU, Y. and WANG, L. (2021). Resampling-based confidence intervals for model-free robust inference on optimal treatment regimes. *Biometrics* **77** 465–476. MR4307648 https://doi.org/10.1111/biom.13337
YANG, Y., LUO, R., LI, M., ZHOU, M., ZHANG, W. and WANG, J. (2018). Mean field multiagent reinforcement learning. ArXiv preprint. Available at arXiv:1802.05438.
YAO, L., CHU, Z., LI, S., LI, Y., GAO, J. and ZHANG, A. (2022). A survey on causal inference. Available at arXiv:2002.02770.
ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x
ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100** 681–694. MR3094445 https://doi.org/10.1093/biomet/ast014
ZHANG, D., SUN, L., LI, B., CHEN, C., PAN, G., LI, S. and WU, Z. (2014). Understanding taxi service strategies from taxi gps traces. *IEEE Trans. Intell. Transp. Syst.* **16** 123–135.
ZHANG, K., YANG, Z. and BAŞAR, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control. Stud. Syst. Decis. Control* **325** 321–384. Springer, Cham. MR4328446 https://doi.org/10.1007/978-3-030-60990-0_12
ZHANG, Y., LABER, E. B., DAVIDIAN, M. and TSIATIS, A. A. (2018). Estimation of optimal treatment regimes using lists. *J. Amer. Statist. Assoc.* **113** 1541–1549. MR3902228 https://doi.org/10.1080/01621459.2017.1345743
ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895–904. MR3436715 https://doi.org/10.1111/biom.12354
ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674
ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488
ZHOU, W., ZHU, R. and QU, A. (2021). Estimating optimal infinite horizon dynamic treatment regimes via pt-learning. ArXiv preprint. Available at arXiv:2110.10719.
ZHU, R., ZHAO, Y.-Q., CHEN, G., MA, S. and ZHAO, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* **73** 391–400. MR3665956 https://doi.org/10.1111/biom.12593

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2723–2758
https://doi.org/10.1214/22-AOAS1711

MODELING RACIAL/ETHNIC DIFFERENCES IN COVID-19 INCIDENCE WITH COVARIATES SUBJECT TO NONRANDOM MISSINGNESS

BY ROB TRANGUCCI[1,a], YANG CHEN[1,b] AND JON ZELNER[2,c]

[1]*Department of Statistics, University of Michigan,* [a]*trangucc@umich.edu,* [b]*ychenang@umich.edu*
[2]*Department of Epidemiology & Center for Social Epidemiology and Population Health, University of Michigan School of Public Health,* [c]*jzelner@umich.edu*

Characterizing the cumulative burden of COVID-19 by race/ethnicity is of the utmost importance for public health researchers and policy makers in order to design effective mitigation measures. This analysis is hampered, however, by surveillance case data with substantial missingness in race and ethnicity covariates. Worse yet, this missingness likely depends on the values of these missing covariates; that is, they are not-missing-at-random (NMAR). We propose a Bayesian parametric model that leverages joint information on spatial variation in the disease and covariate missingness processes and can accommodate both MAR and NMAR missingness. We show that the model is locally identifiable when the spatial distribution of the population covariates is known and observed cases can be associated with a spatial unit of observation. We also use a simulation study to investigate the model's finite-sample performance. We compare our model's performance on NMAR data against complete-case analysis and multiple imputation (MI), both of which are commonly used by public health researchers when confronted with missing categorical covariates. Finally, we model spatial variation in cumulative COVID-19 incidence in Wayne County, Michigan, using data from the Michigan Department of Health and Human Services. The analysis suggests that population relative risk estimates by race during the early part of the COVID-19 pandemic in Michigan were understated for non-white residents, compared to white residents, when cases missing race were dropped or had these values imputed using MI.

REFERENCES

AGUAYO, G. A., SCHRITZ, A., RUIZ-CASTELL, M., VILLARROEL, L., VALDIVIA, G., FAGHERAZZI, G., WITTE, D. R. and LAWSON, A. (2020). Identifying hotspots of cardiometabolic outcomes based on a Bayesian approach: The example of Chile. *PLoS ONE* **15**. https://doi.org/10.1371/journal.pone.0235009

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T. P. A., QUARTAGNO, M., CARPENTER, J., VAN BUUREN, S. and RESCHE-RIGON, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statist. Sci.* **33** 160–183. MR3797708 https://doi.org/10.1214/18-STS646

BAKER, J., WHITE, N. and MENGERSEN, K. (2014). Missing in space: An evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. *Int. J. Health Geogr.* **13** 47. https://doi.org/10.1186/1476-072X-13-47

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

BETANCOURT, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434 [stat].

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

BÜRKNER, P.-C., GABRY, J., KAY, M. and VEHTARI, A. (2021). posterior: Tools for working with posterior distributions. R package version 1.0.1.

CLARK, S. J. and HOULE, B. (2014). Validation, replication, and sensitivity testing of Heckman-type selection models to adjust estimates of HIV prevalence. *PLoS ONE* **9** e112563. https://doi.org/10.1371/journal.pone.0112563

Key words and phrases. Missing data, NMAR, surveillance data, categorical covariates.

COOK, S. R., GELMAN, A. and RUBIN, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Statist.* **15** 675–692. MR2291268 https://doi.org/10.1198/106186006X136976

DIGGLE, P. and KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **43** 49–73.

EEKHOUT, I., DE BOER, R. M., TWISK, J. W. R., DE VET, H. C. W. and HEYMANS, M. W. (2012). Missing data: A systematic review of how they are reported and handled. *Epidemiology* **23** 729–732. https://doi.org/10.1097/EDE.0b013e3182576cdb

ELLIOTT, M. N., MORRISON, P. A., FREMONT, A., MCCAFFREY, D. F., PANTOJA, P. and LURIE, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv. Outcomes Res. Methodol.* **9** 69–83. https://doi.org/10.1007/s10742-009-0047-1

FROME, E. L. (1983). The analysis of rates using Poisson regression models. *Biometrics* **39**. https://doi.org/10.2307/2531094

FROME, E. L. and CHECKOWAY, H. (1985). Use of Poisson regression models in estimating incidence rates and ratios. *Amer. J. Epidemiol.* **121**. https://doi.org/10.1093/oxfordjournals.aje.a114001

GABRY, J. and ČEŠNOVAR, R. (2021). cmdstanr: R Interface to 'CmdStan'. Available at https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

GABRY, J. and MAHR, T. (2021). bayesplot: Plotting for Bayesian models. R package version 1.8.1.

GABRY, J., SIMPSON, D., VEHTARI, A., BETANCOURT, M. and GELMAN, A. (2019). Visualization in Bayesian workflow. *J. Roy. Statist. Soc. Ser. A* **182** 389–402. MR3902665 https://doi.org/10.1111/rssa.12378

GAO, Y., KENNEDY, L., SIMPSON, D. and GELMAN, A. (2021). Improving multilevel regression and poststratification with structured priors. *Bayesian Anal.* **16** 719–744. MR4303866 https://doi.org/10.1214/20-BA1223

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

GELMAN, A., SIMPSON, D. and BETANCOURT, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* **19**. https://doi.org/10.3390/e19100555

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

GELMAN, A., VEHTARI, A., SIMPSON, D., MARGOSSIAN, C. C., CARPENTER, B., YAO, Y., KENNEDY, L., GABRY, J., BÜRKNER, P.-C. et al. (2020). Bayesian workflow. arXiv preprint arXiv:2011.01808.

GÓMEZ-RUBIO, V., CAMELETTI, M. and BLANGIARDO, M. (2019). Missing data analysis and imputation via latent Gaussian Markov random fields. arXiv preprint arXiv:1912.10981.

GOVERNOR WHITMER EXECUTIVE ORDER (2020). Executive order 2020-55: Michigan coronavirus task force on racial disparities. Available at https://www.michigan.gov/whitmer/0,9309,7-387-90499_90705-526476--,00.html. Accessed: 2022-02-10.

GUSTAFSON, P. (2015). *Bayesian Inference for Partially Identified Models*: *Exploring the Limits of Limited Data*. *Monographs on Statistics and Applied Probability* **141**. CRC Press, Boca Raton, FL. MR3642458

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

HELD, L., HENS, N., O'NEILL, P. D. and WALLINGA, J. (2019). *Handbook of Infectious Disease Data Analysis*. CRC Press.

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

HOLLAND, R. C., JONES, G. and BENSCHOP, J. (2015). Spatio–temporal modelling of disease incidence with missing covariate values. *Epidemiol. Infect.* **143**. https://doi.org/10.1017/S0950268814002854

KEELING, M. J. and ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ. MR2354763

KENNEDY, L., KHANNA, K., SIMPSON, D. and GELMAN, A. (2020). Using sex and gender in survey adjustment. arXiv preprint arXiv:2009.14401.

LABGOLD, K., HAMID, S., SHAH, S., GANDHI, N. R., CHAMBERLAIN, A., KHAN, F., KHAN, S., SMITH, S., WILLIAMS, S. et al. (2021). Estimating the unknown: Greater racial and ethnic disparities in COVID-19 burden after accounting for missing race and ethnicity data. *Epidemiology* **32** 157–161. https://doi.org/10.1097/EDE.0000000000001314

LASH, T. L., VANDERWEELE, T. J., HANEUSE, S. and ROTHMAN, K. J. (2021). *Modern Epidemiology*, 4th ed. Lippincott Williams & Wilkins.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

LI, Y., BROWN, P., GESINK, D. C. and RUE, H. (2012). Log Gaussian Cox processes and spatially aggregated disease incidence data. *Stat. Methods Med. Res.* **21** 479–507. MR3190624 https://doi.org/10.1177/0962280212446326

LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90** 1112–1121. MR1354029

LITTLE, R. (2009). Selection and pattern-mixture models. In *Longitudinal Data Analysis. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 409–431. CRC Press, Boca Raton, FL. MR1500128

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LIUBLINSKA, V. and RUBIN, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat. Med.* **33** 4170–4185. MR3267402 https://doi.org/10.1002/sim.6197

MANSON, S., SCHROEDER, J., VAN RIPER, D., KUGLER, T. and RUGGLES, S. (2021). IPUMS national historical geographic information system: Version 16.0 [dataset]. IPUMS, Minneapolis, MN. https://doi.org/10.18128/D050.V16.0

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

MILLETT, G. A., JONES, A. T., BENKESER, D., BARAL, S., MERCER, L., BEYRER, C., HONERMANN, B., LANKIEWICZ, E., MENA, L. et al. (2020). Assessing differential impacts of COVID-19 on black communities. *Ann. Epidemiol.* **47** 37–44. https://doi.org/10.1016/j.annepidem.2020.05.003

MUKERJEE, R. and SUTRADHAR, B. C. (2002). On the positive definiteness of the information matrix under the binary and Poisson mixed models. *Ann. Inst. Statist. Math.* **54** 355–366. MR1910178 https://doi.org/10.1023/A:1022478119885

MICHIGAN DEPARTMENT OF HEALTH AND HUMAN SERVICES (2020). Michigan state and local public health COVID-19 standard operating procedures 41. Michigan Dept. Health and Human Services, Lansing, MI.

PERKINS, N. J., COLE, S. R., HAREL, O., TCHETGEN TCHETGEN, E. J., SUN, B., MITCHELL, E. M. and SCHISTERMAN, E. F. (2018). Principled approaches to missing data in epidemiologic studies. *Amer. J. Epidemiol.* **187** 568–575. https://doi.org/10.1093/aje/kwx348

PRESS OFFICE OF OFFICE OF MICHIGAN GOVERNOR (2020). Governor Whitmer creates the Michigan coronavirus task force on racial disparities. Available at https://www.michigan.gov/coronavirus/0,9753,7-406-98163-525224--,00.html. Accessed: 2022-02-10.

R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAO, C. R. (2002). *Linear Statistical Inference and Its Applications*, 2nd ed., paperback ed. Wiley. MR0346957

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591. MR0436944 https://doi.org/10.2307/1913267

ROY, J. and DANIELS, M. J. (2008). A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics* **64** 538–545, 668. MR2432424 https://doi.org/10.1111/j.1541-0420.2007.00884.x

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196 https://doi.org/10.1093/biomet/63.3.581

SIDI, Y. and HAREL, O. (2018). The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Soc. Sci. Med.* **209** 169–173. https://doi.org/10.1016/j.socscimed.2018.05.037

SIMPSON, D., ILLIAN, J. B., LINDGREN, F., SØRBYE, S. H. and RUE, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika* **103** 49–70. MR3465821 https://doi.org/10.1093/biomet/asv064

STASNY, E. A. (1991). Hierarchical models for the probabilities of a survey classification and nonresponse: An example from the national crime survey. *J. Amer. Statist. Assoc.* **86** 296–303. https://doi.org/10.1080/01621459.1991.10475033

STAVSETH, M. R., CLAUSEN, T. and RØISLIEN, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Med.* **7**. https://doi.org/10.1177/2050312118822912

STAN DEVELOPMENT TEAM (2021). Stan modeling language users guide and reference manual, v2.27.

TRANGUCCI, R., CHEN, Y. and ZELNER, J. (2023). Supplement to "Modeling racial/ethnic differences in COVID-19 incidence with covariates subject to nonrandom missingness." https://doi.org/10.1214/22-AOAS1711SUPPA, https://doi.org/10.1214/22-AOAS1711SUPPB

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

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. D. O'Neill and J. Wallinga, eds.) 455–475 CRC Press. Chapter 23.

WATANABE, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. *Cambridge Monographs on Applied and Computational Mathematics* **25**. Cambridge Univ. Press, Cambridge. MR2554932 https://doi.org/10.1017/CBO9780511800474

ZANGENEH, S. Z. (2012). Model-based methods for robust finite population inference in the presence of external information. Ph.D. thesis, Univ. Michigan.

ZELNER, J., TRANGUCCI, R., NARAHARISETTI, R., CAO, A., MALOSH, R., BROEN, K., MASTERS, N. and DELAMATER, P. (2021). Racial disparities in coronavirus disease 2019 (COVID-19) mortality are driven by unequal infection risks. *Clin*. *Infect*. *Dis*. **72** e88–e95. https://doi.org/10.1093/cid/ciaa1723

ZHANG, G., ROSE, C. E., ZHANG, Y., LI, R., LEE, F. C., MASSETTI, G. and ADAMS, L. E. (2022). Multiple imputation of missing race and ethnicity in CDC COVID-19 case-level surveillance data. *Int*. *J*. *Stat*. *Med*. *Res*. **11**. https://doi.org/10.6000/1929-6029.2022.11.01

ZHOU, X. and REITER, J. P. (2010). A note on Bayesian inference after multiple imputation. *Amer. Statist*. **64** 159–163. MR2757007 https://doi.org/10.1198/tast.2010.09109

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2759–2795
https://doi.org/10.1214/22-AOAS1722

MODEL-INDEPENDENT DETECTION OF NEW PHYSICS SIGNALS USING INTERPRETABLE SEMISUPERVISED CLASSIFIER TESTS

BY PURVASHA CHAKRAVARTI[1,a], MIKAEL KUUSELA[2,b], JING LEI[3,d] AND LARRY WASSERMAN[2,c]

[1]*Department of Statistical Science, University College London,* [a]*p.chakravarti@ucl.ac.uk*

[2]*Department of Statistics & Data Science, and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University,* [b]*mkuusela@andrew.cmu.edu,* [c]*larry@stat.cmu.edu*

[3]*Department of Statistics & Data Science, Carnegie Mellon University,* [d]*jinglei@stat.cmu.edu*

A central goal in experimental high energy physics is to detect new physics signals that are not explained by known physics. In this paper we aim to search for new signals that appear as deviations from known Standard Model physics in high-dimensional particle physics data. To do this, we determine whether there is any statistically significant difference between the distribution of Standard Model background samples and the distribution of the experimental observations which are a mixture of the background and a potential new signal. Traditionally, one also assumes access to a sample from a model for the hypothesized signal distribution. Here we instead investigate a model-independent method that does not make any assumptions about the signal and uses a semisupervised classifier to detect the presence of the signal in the experimental data. We construct three test statistics using the classifier: an estimated likelihood ratio test (LRT) statistic, a test based on the area under the ROC curve (AUC), and a test based on the misclassification error (MCE). Additionally, we propose a method for estimating the signal strength parameter and explore active subspace methods to interpret the proposed semisupervised classifier in order to understand the properties of the detected signal. We also propose a score test statistic that can be used in the model-dependent setting. We investigate the performance of the methods on a simulated data set related to the search for the Higgs boson at the Large Hadron Collider at CERN. We demonstrate that the semisupervised tests have power competitive with the classical supervised methods for a well-specified signal but much higher power for an unexpected signal which might be entirely missed by the supervised tests.

REFERENCES

ADAM-BOURDARIOS, C., COWAN, G., GERMAIN, C., GUYON, I., KÉGL, B. and ROUSSEAU, D. (2015). The Higgs boson machine learning challenge. In *Proceedings of the NIPS* 2014 *Workshop on High-Energy Physics and Machine Learning* (G. Cowan, C. Germain, I. Guyon, B. Kégl and D. Rousseau, eds.). *Proceedings of Machine Learning Research* **42** 19–55. PMLR, Montreal, Canada.

ANDREASSEN, A., NACHMAN, B. and SHIH, D. (2020). Simulation assisted likelihood-free anomaly detection. *Phys. Rev. D* **101** 095004.

ATLAS COLLABORATION (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716** 1–29.

ATLAS COLLABORATION (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal.

ATLAS COLLABORATION (2019). A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. *Eur. Phys. J. C* **79** 120.

ATLAS COLLABORATION AND CMS COLLABORATION (2011). LHC Higgs Combination Group, Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report, CMS-NOTE-2011-005.

Key words and phrases. Collective anomaly detection, active subspace, mixture proportion estimation, signal strength estimation, likelihood ratio test, high-dimensional two-sample testing, Large Hadron Collider.

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

BEHNKE, O., KRÖNINGER, K., SCHOTT, G. and SCHÖRNER-SADENIUS, T. (2013). *Data Analysis in High Energy Physics*: *A Practical Guide to Statistical Methods*. Wiley, New York.

BHAT, P. C. (2011). Multivariate analysis methods in particle physics. *Annu. Rev. Nucl. Part. Sci.* **61** 281–309.

BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. and LINDSAY, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.* **46** 373–388.

BOSTRÖM, H. (2008). Calibrating random forests. In 2008 *Seventh International Conference on Machine Learning and Applications* 121–126.

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

CASA, A. and MENARDI, G. (2018). Nonparametric semisupervised classification for signal detection in high energy physics. ArXiv preprint. Available at arXiv:1809.02977.

CDF COLLABORATION (2008). Model-independent and quasi-model-independent search for new physics at CDF. *Phys. Rev. D* **78** 012002.

CDF COLLABORATION (2009). Global search for new physics with $2.0\,fb^{-1}$ at CDF. *Phys. Rev. D* **79** 011101.

CHAKRAVARTI, P., KUUSELA, M., LEI, J. and WASSERMAN, L. (2023). Supplement to "Model-independent detection of signals using interpretable semi-supervised classifier tests." https://doi.org/10.1214/22-AOAS1722SUPPA, https://doi.org/10.1214/22-AOAS1722SUPPB

CHANDOLA, V., BANERJEE, A. and KUMAR, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.* **41** 1–58.

CHOUDALAKIS, G. (2008). Model independent search for new physics at the Tevatron. ArXiv preprint. Available at arXiv:0805.3954.

CMS COLLABORATION (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716** 30–61.

CMS COLLABORATION (2017). MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV. CMS Physics Analysis Summary CMS-PAS-EXO-14/016.

CMS COLLABORATION (2020). MUSiC, a model unspecific search for new physics, in pp collisions at sqrt (s) = 13 TeV. Technical Report, Technical report CMS-PAS-EXO-19-008, CERN, Geneva.

COLLINS, J., HOWE, K. and NACHMAN, B. (2018). Anomaly detection for resonant new physics with machine learning. *Phys. Rev. Lett.* **121** 241803. https://doi.org/10.1103/PhysRevLett.121.241803

COLLINS, J. H., HOWE, K. and NACHMAN, B. (2019). Extending the search for new resonances with machine learning. *Phys. Rev. D* **99** 014038.

CONSTANTINE, P. G. (2015). *Active Subspaces*: *Emerging Ideas for Dimension Reduction in Parameter Studies*. *SIAM Spotlights* **2**. SIAM, Philadelphia, PA. MR3486165 https://doi.org/10.1137/1.9781611973860

CONSTANTINE, P. G., DOW, E. and WANG, Q. (2014). Active subspace methods in theory and practice: Applications to Kriging surfaces. *SIAM J. Sci. Comput.* **36** A1500–A1524. MR3233940 https://doi.org/10.1137/130916138

CONSTANTINE, P. G., EMORY, M., LARSSON, J. and IACCARINO, G. (2015). Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet. *J. Comput. Phys.* **302** 1–20. MR3404505 https://doi.org/10.1016/j.jcp.2015.09.001

COWAN, G., CRANMER, K., GROSS, E. and VITELLS, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C* **71** 1554.

CRANMER, K. (2015). Practical statistics for the LHC. ArXiv preprint. Available at arXiv:1503.07622.

CRANMER, K., PAVEZ, J. and LOUPPE, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. ArXiv preprint. Available at arXiv:1506.02169.

CUI, C., ZHANG, K., DAULBAEV, T., GUSAK, J., OSELEDETS, I. and ZHANG, Z. (2020). Active subspace of neural networks: Structural analysis and universal attacks. *SIAM J. Math. Data Sci.* **2** 1096–1122. MR4168214 https://doi.org/10.1137/19M1296070

D'AGNOLO, R. T. and WULZER, A. (2019). Learning new physics from a machine. *Phys. Rev. D* **99** 015014.

D'AGNOLO, R. T., GROSSO, G., PIERINI, M., WULZER, A. and ZANETTI, M. (2021). Learning multivariate new physics. *Eur. Phys. J. C* **81** 1–21.

D'AGNOLO, R. T., GROSSO, G., PIERINI, M., WULZER, A. and ZANETTI, M. (2022). Learning new physics from an imperfect machine. *Eur. Phys. J. C* **82** 1–37.

D0 COLLABORATION (2012). Model independent search for new phenomena in pp (bar) collisions at $\sqrt{s} =$ 1.96 TeV. *Phys. Rev. D* **85**.

DAUNCEY, P., KENZIE, M., WARDLE, N. and DAVIES, G. (2015). Handling uncertainties in background shapes: The discrete profiling method. *J. Instrum.* **10** P04015.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

DOBSON, A. J. and BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*, 4th ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3890007

DORIGO, T. and DE CASTRO, P. (2020). Dealing with nuisance parameters using machine learning in high energy physics: A review. ArXiv preprint. Available at arXiv:2007.09121.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

FRIEDMAN, J. H. (2003). On multivariate goodness-of-fit and two-sample testing. In *PHYSTAT* 2003, *SLAC*, *Stanford*, *California*.

GHOSH, A., NACHMAN, B. and WHITESON, D. (2021). Uncertainty-aware machine learning for high energy physics. *Phys*. *Rev. D* **104** 056026.

GHOSH, J. K. and SEN, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, *Vol*. *II* (*Berkeley*, *Calif*., 1983). *Wadsworth Statist./Probab. Ser.* 789–806. Wadsworth, Belmont, CA. MR0822065

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.

GRÖMPING, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *Amer. Statist*. **63** 308–319. MR2751747 https://doi.org/10.1198/tast.2009.08199

H1 COLLABORATION (2004). A general search for new phenomena in ep scattering at HERA. *Phys. Lett. B* **602** 14–30.

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HANLEY, J. A. et al. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging* **29** 307–335.

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

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann*. *Statist*. **16** 772–783. MR0947577 https://doi.org/10.1214/aos/1176350835

ISHWARAN, H. (2007). Variable importance in binary regression trees and forests. *Electron*. *J. Stat*. **1** 519–537. MR2357716 https://doi.org/10.1214/07-EJS039

KASIECZKA, G., NACHMAN, B., SHIH, D., AMRAM, O., ANDREASSEN, A., BENKENDORFER, K., BORTOLATO, B., BROOIJMANS, G., CANELLI, F. et al. (2021). The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics. ArXiv preprint. Available at arXiv:2101.08320.

KIM, I., LEE, A. B. and LEI, J. (2019). Global and local two-sample tests via regression. *Electron*. *J. Stat*. **13** 5253–5305. MR4043073 https://doi.org/10.1214/19-EJS1648

KIM, I., RAMDAS, A., SINGH, A. and WASSERMAN, L. (2021). Classification accuracy as a proxy for two-sample testing. *Ann*. *Statist*. **49** 411–434. MR4206684 https://doi.org/10.1214/20-AOS1962

KNUTESON, B. (2000). Ph.D. thesis, University of California at Berkeley.

KUUSELA, M., VATANEN, T., MALMI, E., RAIKO, T., AALTONEN, T. and NAGAI, Y. (2012). Semi-supervised anomaly detection–towards model-independent searches of new physics. In *Journal of Physics*: *Conference Series* **368** 012032. IOP Publishing, Bristol.

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

LYONS, L. and WARDLE, N. (2018). Statistical issues in searches for new phenomena in high energy physics. *J. Phys*. *G*, *Nucl*. *Part*. *Phys*. **45** 033001.

METZ, C. E. (1978). Basic principles of ROC analysis. In *Seminars in Nuclear Medicine* **8** 283–298. Elsevier, Amsterdam.

NACHMAN, B. (2020). A guide for deploying deep learning in LHC searches: How to achieve optimality and account for uncertainty. *SciPost Phys*. **8** Paper No. 090. MR4196320 https://doi.org/10.21468/scipostphys.8.6.090

NACHMAN, B. and SHIH, D. (2020). Anomaly detection with density estimation. *Phys*. *Rev. D* **101** 075042.

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *J. R. Stat*. *Soc*., *A* **135** 370–384.

NEWCOMBE, R. G. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: Asymptotic methods and evaluation. *Stat. Med.* **25** 559–573. MR2222114 https://doi.org/10.1002/sim.2324

NICULESCU-MIZIL, A. and CARUANA, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the* 22*nd International Conference on Machine Learning*. *ICML* 2005 625–632. Association for Computing Machinery, New York, NY, USA.

PARTICLE DATA GROUP (2020). Review of particle physics. *PTEP* **2020** 083C01.

PEYRÉ, G., CUTURI, M. et al. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607.

RADOVIC, A., WILLIAMS, M., ROUSSEAU, D., KAGAN, M., BONACORSI, D., HIMMEL, A., AURISANO, A., TERAO, K. and WONGJIRAD, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560** 41–48. https://doi.org/10.1038/s41586-018-0361-2

REISS, R.-D. (1993). *A Course on Point Processes*. *Springer Series in Statistics*. Springer, New York. MR1199815 https://doi.org/10.1007/978-1-4613-9308-5

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

SHRIKUMAR, A., GREENSIDE, P. and KUNDAJE, A. (2017). Learning important features through propagating activation differences. ArXiv preprint. Available at arXiv:1704.02685.

SOHA, A. L. (2008). General searches for new physics. In 34*th International Conference on High Energy Physics*.

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

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

SUNDARARAJAN, M., TALY, A. and YAN, Q. (2017). Axiomatic attribution for deep networks. ArXiv preprint. Available at arXiv:1703.01365.

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

VATANEN, T., KUUSELA, M., MALMI, E., RAIKO, T., AALTONEN, T. and NAGAI, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *The* 2012 *International Joint Conference on Neural Networks* (*IJCNN*) 1–8. IEEE, New York.

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. ArXiv preprint. Available at arXiv:2004.03683.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2796–2820
https://doi.org/10.1214/23-AOAS1737

PREDICTIVE INFERENCE FOR TRAVEL TIME ON TRANSPORTATION NETWORKS

BY MOHAMAD ELMASRI[1,a], AURÉLIE LABBE[2,b], DENIS LAROCQUE[2,c] AND LAURENT CHARLIN[2,d]

[1]*Department of Statistical Sciences, University of Toronto,* [a]*mohamad.elmasri@utoronto.ca*
[2]*Department of Decision Sciences, HEC Montréal,* [b]*aurelie.labbe@hec.ca,* [c]*denis.larocque@hec.ca,* [d]*laurent.charlin@hec.ca*

Recent statistical methods fitted on large-scale GPS data can provide accurate estimations of the expected travel time between two points. However, little is known about the distribution of travel time, which is key to decision-making across a number of logistic problems. With sufficient data single road-segment travel time can be well approximated. The challenge lies in understanding how to aggregate such information over a route to arrive at the route-distribution of travel time. We develop a novel statistical approach to this problem. We show that, under general conditions and without assuming a distribution of speed, travel time divided by route distance follows a Gaussian distribution with route-invariant population mean and variance. We develop efficient inference methods for these parameters and propose asymptotically tight population prediction intervals for travel time. Using traffic flow information, we further develop a trip-specific Gaussian-based predictive distribution, resulting in tight prediction intervals for short and long trips. Our methods, implemented in an R-package,[1] are illustrated in a real-world case study using mobile GPS data, showing that our trip-specific and population intervals both achieve the 95% theoretical coverage levels. Compared to alternative approaches, our trip-specific predictive distribution achieves: (a) the theoretical coverage at every level of significance, (b) tighter prediction intervals, (c) less predictive bias, and (d) more efficient estimation and prediction procedures. This makes our approach promising for low-latency, large-scale transportation applications.

REFERENCES

ALDOUS, D. (1991). Applications of random walks on finite graphs. In *Selected Proceedings of the Sheffield Symposium on Applied Probability* (*Sheffield*, 1989). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **18** 12–26. IMS, Hayward, CA. MR1193058 https://doi.org/10.1214/lnms/1215459284

BARRAT, A., BARTHÉLEMY, M. and VESPIGNANI, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge Univ. Press, Cambridge. MR2797803 https://doi.org/10.1017/CBO9780511791383

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

BRITTON, T. and O'NEILL, P. D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Stat.* **29** 375–390. MR1925565 https://doi.org/10.1111/1467-9469.00296

BUDGE, S., INGOLFSSON, A. and ZEROM, D. (2010). Empirical analysis of ambulance travel times: The case of Calgary emergency medical services. *Manage. Sci.* **56** 716–723.

BURK, W. J., STEGLICH, C. E. and SNIJDERS, T. A. (2007). Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. *Int. J. Behav. Dev.* **31** 397–404.

CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference* **2**. Wadsworth, Belmont.

EINSIEDLER, M. and WARD, T. (2013). *Ergodic Theory*. Springer, Berlin.

ELMASRI, M., LABBE, A., LAROCQUE, D. and CHARLIN, L. (2023). Supplement to "Predictive inference for travel time on transportation networks." https://doi.org/10.1214/23-AOAS1737SUPPA, https://doi.org/10.1214/23-AOAS1737SUPPB

Key words and phrases. Central limit theorem, mixing sequence, dynamical systems, prediction intervals, distribution of travel time.

[1]Available at https://github.com/melmasri/traveltimeCLT.

GARDNER, W. A., NAPOLITANO, A. and PAURA, L. (2006). Cyclostationarity: Half a century of research. *Signal Process*. **86** 639–697.

GEISSER, S. (1993). *Predictive Inference*: *An Introduction*. *Monographs on Statistics and Applied Probability* **55**. CRC Press, New York. . MR1252174 https://doi.org/10.1007/978-1-4899-4467-2

GEROLIMINIS, N. and DAGANZO, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transp. Res*., *Part B*, *Methodol*. **42** 759–770.

GEROLIMINIS, N. and SKABARDONIS, A. (2006). Real time vehicle reidentification and performance measures on signalized arterials. In 9*th International IEEE Conference on Intelligent Transportation Systems*, *Toronto*, *Canada* 188–193.

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

GUO, F., LI, Q. and RAKHA, H. (2012). Multistate travel time reliability models with skewed component distributions. *Transp. Res. Rec*. **2315** 47–53.

HERRNDORF, N. (1983). The invariance principle for φ-mixing sequences. *Z. Wahrsch. Verw. Gebiete* **63** 97–108. MR0699789 https://doi.org/10.1007/BF00534180

HUNTER, T., DAS, T., ZAHARIA, M., ABBEEL, P. and BAYEN, A. M. (2013). Large-scale estimation in cyber-physical systems using streaming data: A case study with arterial traffic estimation. *IEEE Trans. Autom. Sci. Eng*. **10** 884–898.

HUNTER, T., HERRING, R., ABBEEL, P. and BAYEN, A. (2009). Path and travel time inference from GPS probe vehicle data. *NIPS Analyzing Networks and Learning with Graphs* **12** 1–8.

JENELIUS, E. and KOUTSOPOULOS, H. N. (2013). Travel time estimation for urban road networks using low frequency probe vehicle data. *Transp. Res*., *Part B*, *Methodol*. **53** 64–81.

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

KOLACZYK, E. D. and CSÁRDI, G. (2014). *Statistical Analysis of Network Data with R. Use R!* Springer, New York. MR3288852 https://doi.org/10.1007/978-1-4939-0983-4

LI, C., PARKER, D. and HAO, Q. (2021). Vehicle dispatch in on-demand ride-sharing with stochastic travel times. In 2021 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), *Prague*, *Czech Republic* 5966–5972.

LI, M., JIANG, G. and LO, H. K. (2022). Pricing strategy of ride-sourcing services under travel time variability. *Transp. Res*., *Part E*, *Logist. Transp. Rev.* **159** 102631.

LI, X., GAO, J., WANG, C., HUANG, X. and NIE, Y. (2022). Ride-sharing matching under travel time uncertainty through data-driven robust optimization. *IEEE Access* **10** 116931–116941.

LONG, J., TAN, W., SZETO, W. and LI, Y. (2018). Ride-sharing with travel time uncertainty. *Transp. Res*., *Part B*, *Methodol*. **118** 143–171.

MA, Z., KOUTSOPOULOS, H. N., FERREIRA, L. and MESBAH, M. (2017). Estimation of trip travel time distribution using a generalized Markov chain approach. *Transp. Res*., *Part C*, *Emerg. Technol*. **74** 1–21.

NEWSON, P. and KRUMM, J. (2009). Hidden Markov map matching through noise and sparseness. In *Proceedings of the* 17*th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 336–343. ACM, New York.

PELIGARD, M. and SURESH, R. (1995). Estimation of variance of partial sums of an associated sequence of random variables. *Stochastic Process. Appl*. **56** 307–319. MR1325225 https://doi.org/10.1016/0304-4149(94)00065-2

PELIGRAD, M. (1996). On the asymptotic normality of sequences of weak dependent random variables. *J. Theoret. Probab*. **9** 703–715. MR1400595 https://doi.org/10.1007/BF02214083

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **69** 741–796. MR2368570 https://doi.org/10.1111/j.1467-9868.2007.00610.x

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

SNIJDERS, T., STEGLICH, C. and SCHWEINBERGER, M. (2017). Modeling the coevolution of networks and behavior. In *Longitudinal Models in the Behavioral and Related Sciences* 41–71. Routledge, London.

TREIBER, M., HENNECKE, A. and HELBING, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* **62** 1805.

WANG, H., TANG, X., KUO, Y.-H., KIFER, D. and LI, Z. (2019). A simple baseline for travel time estimation using large-scale trip data. *ACM Trans. Intell. Syst. Technol*. **10** 19.

WESTGATE, B. S., WOODARD, D. B., MATTESON, D. S. and HENDERSON, S. G. (2013). Travel time estimation for ambulances using Bayesian data augmentation. *Ann. Appl. Stat*. **7** 1139–1161. MR3113504 https://doi.org/10.1214/13-AOAS626

WESTGATE, B. S., WOODARD, D. B., MATTESON, D. S. and HENDERSON, S. G. (2016). Large-network travel time distribution estimation for ambulances. *European J. Oper. Res.* **252** 322–333.

WILLIAMS, B. M. and HOEL, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *J. Transp. Eng.* **129** 664–672.

WOODARD, D., NOGIN, G., KOCH, P., RACZ, D., GOLDSZMIDT, M. and HORVITZ, E. (2017). Predicting travel time reliability using mobile phone GPS data. *Transp. Res., Part C, Emerg. Technol.* **75** 30–44.

WU, W. B. (2009). Recursive estimation of time-average variance constants. *Ann. Appl. Probab.* **19** 1529–1552. MR2538079 https://doi.org/10.1214/08-AAP587

ZHANG, K., JIA, N., ZHENG, L. and LIU, Z. (2019). A novel generative adversarial network for estimation of trip travel time distribution with trajectory data. *Transp. Res., Part C, Emerg. Technol.* **108** 223–244.

ZHENG, F. and VAN ZUYLEN, H. J. (2013). Urban link travel time estimation based on sparse probe vehicle data. *Transp. Res., Part C, Emerg. Technol.* **31** 145–157.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2821–2842
https://doi.org/10.1214/23-AOAS1738

A FRAMEWORK FOR COVARIATE-SPECIFIC ROC CURVE ESTIMATION, WITH APPLICATION TO BIOMETRIC RECOGNITION

BY XIAOCHEN ZHU[1,a], MARTIN SLAWSKI[1,b] AND LIANSHENG TANG[2,c]

[1]*Department of Statistics, George Mason University,* [a]*xzhu11@gmu.edu,* [b]*mslawsk3@gmu.edu*
[2]*Department of Statistics and Data Science, National Center of Forensic Science, University of Central Florida,* [c]*liansheng.tang@ucf.edu*

Biometric traits, such as fingerprints, facial images, and teeth impressions, are often used in forensic analysis to identify crime suspects. Matching such biometric traits is not perfect, and recent reports have indicated the need for quantifiable measures of error rates for (these) possible matches. Often, comparisons between two sets of a trait are scored with a higher score indicating a higher likelihood that the sets are a match. Adjustment of the cutoff for which a match is declared yields a trade-off between false positive and false negative decisions that can be represented by an ROC curve. In this paper we study modeling of such ROC curves conditional on covariates, for example, demographic information about source subjects, quality properties of the underlying biometric measurements, or characteristics of forensic examiners; quantifying how error rates vary in dependence of such covariates is often considerably more meaningful in biometrics and forensics than the "raw" error rates based on the pooled data. We herein develop a framework for estimating covariate-specific ROC curves that integrates robustness, heteroscedasticity, and stochastic ordering. The latter is of specific relevance in the given application since biometric recognition systems are typically calibrated to assign higher scores to matching pairs than to nonmatching pairs. The proposed methodology is demonstrated on accuracy of face recognition and fingerprint matching and also has potential in other domains of application like medical diagnostics.

REFERENCES

ARCONES, M. A., KVAM, P. H. and SAMANIEGO, F. J. (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *J. Amer. Statist. Assoc.* **97** 170–182. MR1947278 https://doi.org/10.1198/016214502753479310

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

CAI, T. and MOSKOWITZ, C. S. (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics* **5** 573–586.

CAO, K., NGUYEN, D.-L., TYMOSZEK, C. and JAIN, A. K. (2019). End-to-end latent fingerprint search. *IEEE Trans. Inform. Forensics Secur.* **15** 880–894.

CHEN, B., LI, P., QIN, J. and YU, T. (2016). Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *J. Amer. Statist. Assoc.* **111** 861–874. MR3538711 https://doi.org/10.1080/01621459.2015.1066681

DODD, L. E. (2001). Regression methods for areas and partial areas under the receiver-operating characteristic curve. Ph.D. Thesis, Univ., Seattle, WA.

DONG, T., TANG, L. L. and ROSENBERGER, W. F. (2014). Optimal sampling ratios in comparative diagnostic trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 499–514. MR3238164 https://doi.org/10.1111/rssc.12043

DUAN, X. and ZHOU, X.-H. (2013). Composite quantile regression for the receiver operating characteristic curve. *Biometrika* **100** 889–900. MR3142339 https://doi.org/10.1093/biomet/ast025

FARAGGI, D. and REISER, B. (2002). Estimation of the area under the ROC curve. *Stat. Med.* **21** 3093–3106.

Key words and phrases. ROC regression, median regression, order constraint, heteroscedastic modeling, facial recognition.

González-Manteiga, W., Pardo-Fernández, J. C. and Van Keilegom, I. (2011). ROC curves in nonparametric location-scale regression models. *Scand. J. Stat.* **38** 169–184. MR2760145 https://doi.org/10.1111/j.1467-9469.2010.00693.x

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Monographs on Statistics and Applied Probability* **58**. CRC Press, London. MR1270012 https://doi.org/10.1007/978-1-4899-4473-3

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

He, X. (1997). Quantile curves without crossing. *Amer. Statist.* **51** 186–192.

Holdren, J., Lander, E., Press, W., Savitz, M., Austin, W., Chyba, C. et al. (2016). Report to the president forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Subcommittee on the Social and Behavioral Sciences Team: United States Government.

Hornak, L., Williams, L., Cukic, B., Ross, A., Morris, K., Dawson, J., Crihalmeanu, S., Kalka, N. and Kayal, N. (2009). FBI biometric collection of people (biocop)—next generation identification—phase 1 (2008–2009). 2008 Biometric Collection Project 08-06-2008 to 12-31-2009 FINAL REPORT.

Hwang, B. S. and Chen, Z. (2015). An integrated Bayesian nonparametric approach for stochastic and variability orders in ROC curve estimation: An application to endometriosis diagnosis. *J. Amer. Statist. Assoc.* **110** 923–934. MR3420673 https://doi.org/10.1080/01621459.2015.1023806

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

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press.

O'Toole, A. J., Phillips, P. J., An, X. and Dunlop, J. (2012). Demographic effects on estimates of automatic face recognition performance. *Image Vis. Comput.* **30** 169–176.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* **84** 595–608. MR1603993 https://doi.org/10.1093/biomet/84.3.595

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56** 352–359. MR1795014 https://doi.org/10.1111/j.0006-341X.2000.00352.x

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series* **28**. Oxford Univ. Press, Oxford. MR2260483

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., Dunlop, J., Lui, Y. M., Sahibzada, H. et al. (2012). The good, the bad, and the ugly face challenge problem. *Image Vis. Comput.* **30** 177–185.

Phillips, P. J., Bowyer, K. W. and Flynn, P. J. (2007). Comment on the casia version 1.0 iris dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**.

Phillips, P. J., Boyer, K. W., Flynn, A. J., O'Toole, P. J., Schott, C. L., Scruggs, W. T. and Sharpe, M. (2007). Face recognition vendor test (FRVT) 2006 and iris challenge evaluation (ICE) 2006 large-scale results. *NIST Interagency/Internal Report*.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. USA* **115** 6171–6176.

Tang, C.-F., Wang, D. and Tebbs, J. M. (2017). Nonparametric goodness-of-fit tests for uniform stochastic ordering. *Ann. Statist.* **45** 2565–2589. MR3737902 https://doi.org/10.1214/16-AOS1535

Tang, L., Emerson, S. S. and Zhou, X.-H. (2008). Nonparametric and semiparametric group sequential methods for comparing accuracy of diagnostic tests. *Biometrics* **64** 1137–1145. MR2522261 https://doi.org/10.1111/j.1541-0420.2008.01000.x

Tang, L., Kang, L., Liu, C., Schisterman, E. F. and Liu, A. (2013). An additive selection of markers to improve diagnostic accuracy based on a discriminatory measure. *Acad. Radiol.* **20** 854–862.

Tang, L. and Liu, A. (2009). Sample size recalculation in sequential diagnostic trials. *Biostatistics* **11** 151–163.

Tang, L. and Zhou, X.-H. (2009). Semiparametric inferential procedures for comparing multivariate ROC curves with interaction terms. *Statist. Sinica* **19** 1203–1221. MR2536152

Ulery, B. T., Hicklin, R. A., Buscaglia, J. and Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proc. Natl. Acad. Sci. USA* **108** 7733–7738.

Ulery, B. T., Hicklin, R. A., Buscaglia, J. and Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *Public Library of Science One* **7** e32800.

Westling, T., Downes, K. J. and Small, D. S. (2023). Nonparametric maximum likelihood estimation under a likelihood ratio order. *Statist. Sinica* **33** 573–591. MR4575315 https://doi.org/10.5705/ss.202020.0207

White, D., Norell, K., Phillips, P. J. and O'Toole, A. J. (2017). Human factors in forensic face identification. In *Handbook of Biometrics for Forensic Science* 195–218. Springer, Berlin.

WOOD, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM J. Sci. Comput.* **15** 1126–1133. MR1289157 https://doi.org/10.1137/0915069

YOON, S. and JAIN, A. K. (2015). Longitudinal study of fingerprint recognition. *Proc. Natl. Acad. Sci. USA* **112** 8555–8560.

YU, T., LI, P. and QIN, J. (2017). Density estimation in the two-sample problem with likelihood ratio ordering. *Biometrika* **104** 141–152. MR3626481 https://doi.org/10.1093/biomet/asw069

YUAN, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **50** 813–829. MR2207010 https://doi.org/10.1016/j.csda.2004.10.008

ZHANG, H., BEVERIDGE, J. R., DRAPER, B. A. and PHILLIPS, P. J. (2015). On the effectiveness of soft biometrics for increasing face verification rates. *Comput. Vis. Image Underst.* **137** 50–62.

ZHANG, W., TANG, L. L., LI, Q., LIU, A. and LEE, M.-L. T. (2020). Order-restricted inference for clustered ROC data with application to fingerprint matching accuracy. *Biometrics* **76** 863–873. MR4151855 https://doi.org/10.1111/biom.13177

ZHOU, X.-H., OBUCHOWSKI, N. A. and MCCLISH, D. K. (2011). *Statistical Methods in Diagnostic Medicine*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2816760 https://doi.org/10.1002/9780470906514

ZHU, X., SLAWSKI, M., PHILLIPS, P. J. and TANG, L. L. (2021). Order-constrained ROC regression with application to facial recognition. *Technometrics* **63** 343–353. MR4296901 https://doi.org/10.1080/00401706.2020.1785549

ZHU, X., SLAWSKI, M. and TANG, L. (2023). Supplement to "A framework for covariate-specific ROC curve estimation, with application to biometric recognition." https://doi.org/10.1214/23-AOAS1738SUPPA, https://doi.org/10.1214/23-AOAS1738SUPPB

ZOU, H. and YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651 https://doi.org/10.1214/07-AOS507

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2843–2864
https://doi.org/10.1214/23-AOAS1740

VARYING IMPACTS OF LETTERS OF RECOMMENDATION ON COLLEGE ADMISSIONS

BY ELI BEN-MICHAEL[1,a], AVI FELLER[2,b] AND JESSE ROTHSTEIN[3,c]

[1]*Department of Statistics and Heinz College, Carnegie Mellon University,* [a]*ebenmichael@cmu.edu*
[2]*Goldman School of Public Policy and Department of Statistics, University of California, Berkeley,* [b]*afeller@berkeley.edu*
[3]*Goldman School of Public Policy and Department of Economics, University of California, Berkeley,* [c]*rothstein@berkeley.edu*

In a pilot program during the 2016–17 admissions cycle, the University of California, Berkeley invited many applicants for freshman admission to submit letters of recommendation. This proved controversial within the university, with concerns that this change would further disadvantage applicants from disadvantaged groups. To inform this debate, we use this pilot as the basis for an observational study of the impact of submitting letters of recommendation on subsequent admission, with the goal of estimating how impacts vary across predefined subgroups. Understanding this variation is challenging in an observational setting because estimated impacts reflect both actual treatment effect variation and differences in covariate balance across groups. To address this, we develop balancing weights that directly optimize for "local balance" within subgroups while maintaining global covariate balance between treated and control units. Applying this approach to the UC Berkeley pilot study yields excellent local and global balance, unlike more traditional weighting methods, which fail to balance covariates within subgroups. We find that the impact of letters of recommendation increases with applicant strength. However, we find little average difference for applicants from disadvantaged groups, although this result is more mixed. In the end we conclude that soliciting letters of recommendation from a broader pool of applicants would not meaningfully change the composition of admitted undergraduates.

REFERENCES

ALVERO, A., GIEBEL, S., GEBRE-MEDHIN, B., ANTONIO, A. L., STEVENS, M. L. and DOMINGUE, B. W. (2021). Essay content is strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. Technical report, Stanford Center for Education Policy Analysis Working Paper.

ANOKE, S. C., NORMAND, S.-L. and ZIGLER, C. M. (2019). Approaches to treatment effect heterogeneity in the presence of confounding. *Stat. Med.* **38** 2797–2815. MR3962143 https://doi.org/10.1002/sim.8143

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

BEN-MICHAEL, E., FELLER, A. and HARTMAN, E. (2021). Multilevel calibration weighting for survey data. Preprint. Available at arXiv:2102.09052.

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2021). The augmented synthetic control method. *J. Amer. Statist. Assoc.* **116** 1789–1803. MR4353714 https://doi.org/10.1080/01621459.2021.1929245

BEN-MICHAEL, E., FELLER, A. and ROTHSTEIN, J. (2023). Supplement to "Varying impacts of letters of recommendation on college admissions." https://doi.org/10.1214/23-AOAS1740SUPP

BEN-MICHAEL, E., HIRSCHBERG, D., FELLER, A. and ZUBIZARRETA, J. (2021). The balancing act for causal inference. Preprint. Available at arXiv:2110.14831.

BICKEL, P. J., HAMMEL, E. A. and O'CONNELL, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* **187** 398–404. https://doi.org/10.1126/science.187.4175.398

BLEEMER, Z. (2022). Affirmative action, mismatch, and economic mobility after California's Proposition 209. *Q. J. Econ.* **137** 115–160.

BOWEN, W. G. and BOK, D. (1996). The shape of the river: Long-term consequences of considering race in college and university admissions. In *The Shape of the River*. Princeton University Press, Princeton.

Key words and phrases. Higher education, observational study, subgroup analysis, balancing weights.

CARVALHO, C., FELLER, A., MURRAY, J., WOODY, S. and YEAGER, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Obs. Stud.* **5** 21–35.

CHALFANT, J. (2017). Letter from Jim Chalfant, Chair of the Academic Senate, to Janet Napolitano. June 20, 2017.

DEVILLE, J. C., SÄRNDAL, C. E. and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *J. Amer. Statist. Assoc.* **88** 1013–1020. https://doi.org/10.1080/01621459.1993.10476369

DONG, J., ZHANG, J. L., ZENG, S. and LI, F. (2020). Subgroup balancing propensity score. *Stat. Methods Med. Res.* **29** 659–676. MR4078241 https://doi.org/10.1177/0962280219870836

FELLER, A. and GELMAN, A. (2015). Hierarchical models for causal effects. In *Emerging Trends in the Social and Behavioral Sciences*: *An Interdisciplinary*, *Searchable*, *and Linkable Resource* 1–16.

GREEN, K. M. and STUART, E. A. (2014). Examining moderation analyses in propensity score methods: Application to depression and substance use. *J. Consult. Clin. Psychol.* **82** 773–783. https://doi.org/10.1037/a0036515

GRIFFIN, B. A., SCHULER, M. S., CEFALU, M., AYER, L., GODLEY, M., GREIFER, N., COFFMAN, D. L. and MCCAFFREY, D. (2022). A tutorial for using propensity score weighting for moderation analysis: An application to smoking disparities among LGB adults. Preprint. Available at arXiv:2204.03345.

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

HAZLETT, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statist. Sinica* **30** 1155–1189. MR4257528 https://doi.org/10.5705/ss.20

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

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

HOUT, M. (2005). Berkeley's comprehensive review method for making freshman admissions decisions: An assessment. Technical report, Univ. California, Berkeley.

KARABEL, J. (2005). *The Chosen*: *The Hidden History of Admission and Exclusion at Harvard*, *Yale*, *and Princeton*. Houghton Mifflin Harcourt, Boston.

KUNCEL, N. R., KOCHEVAR, R. J. and ONES, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *Int. J. Sel. Assess.* **22** 101–107.

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. USA* **116** 4156–4165. https://doi.org/10.1073/pnas.1804597116

LEE, Y., NGUYEN, T. Q. and STUART, E. A. (2021). Partially pooled propensity score models for average treatment effect estimation with multilevel data. *J. Roy. Statist. Soc. Ser. A* **184** 1578–1598. MR4344649 https://doi.org/10.1111/rssa.12741

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

MADERA, J. M., HEBL, M. R. and MARTIN, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *J. Appl. Psychol.* **94** 1591–1599. https://doi.org/10.1037/a0016539

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425. https://doi.org/10.1037/1082-989X.9.4.403

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

ROTHSTEIN, J. M. (2004). College performance predictions and the SAT. *J. Econometrics* **121** 297–317. MR2059876 https://doi.org/10.1016/j.jeconom.2003.10.003

ROTHSTEIN, J. (2017). The impact of letters of recommendation on UC Berkeley admissions in the 2016–17 cycle. Technical report, California Policy Lab.

ROTHSTEIN, J. (2022). Qualitative information in undergraduate admissions: A pilot study of letters of recommendation. *Econ. Educ. Rev.* **89** 102285.

RUBIN, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.

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

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

SCHMADER, T., WHITEHEAD, J. and WYSOCKI, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles* **57** 509–514. https://doi.org/10.1007/s11199-007-9291-4

SORIANO, D., BEN-MICHAEL, E., BICKEL, P., FELLER, A. and PIMENTEL, S. (2020). Interpretable sensitivity analysis for balancing weights. Technical report.

TRIX, F. and PSENKA, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse Soc.* **14** 191–220.

UC BERKELEY (2017). Notice of Meeting: May 11, 2017. University Committee on Affirmative Action, Diversity, and Equity.

UNIVERSITY OF CALIFORNIA BOARD OF REGENTS (2022). Regents Policy 2110: Policy on Augmented Review in Undergraduate Admissions.

VANDERWEELE, T. J. and KNOL, M. J. (2011). Interpretation of subgroup analyses in randomized trials: Heterogeneity versus secondary interventions. *Ann. Intern. Med.* **154** 680–683. https://doi.org/10.7326/0003-4819-154-10-201105170-00008

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WANG, J., WONG, R. K. W., YANG, S. and CHAN, K. C. G. (2022). Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Electron. J. Stat.* **16** 4332–4378. MR4474576 https://doi.org/10.1214/22-ejs2000

YANG, S., LORENZI, E., PAPADOGEORGOU, G., WOJDYLA, D. M., LI, F. and THOMAS, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Stat. Med.* **40** 4294–4309. MR4300087 https://doi.org/10.1002/sim.9029

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672 https://doi.org/10.1080/01621459.2015.1023805

ZUBIZARRETA, J. R. and KEELE, L. (2017). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *J. Amer. Statist. Assoc.* **112** 547–560. MR3671751 https://doi.org/10.1080/01621459.2016.1240683

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2865–2886
https://doi.org/10.1214/23-AOAS1742

BAYESIAN HIERARCHICAL MODELING AND ANALYSIS FOR ACTIGRAPH DATA FROM WEARABLE DEVICES

BY PIERFRANCESCO ALAIMO DI LORO[1,a], MARCO MINGIONE[2,b], JONAH LIPSITT[3,c], CHRISTINA M. BATTEATE[4,e], MICHAEL JERRETT[3,d] AND SUDIPTO BANERJEE[5,f]

[1]*Department GEPLI, LUMSA,* [a]*p.alaimodiloro@lumsa.it*

[2]*Department of Political Sciences, Roma Tre University,* [b]*marco.mingione@uniroma3.it*

[3]*Department of Environmental Health Sciences, University of California, Los Angeles,* [c]*jonahlipsitt@gmail.com,* [d]*mjerrett@ucla.edu*

[4]*Center of Occupational and Environmental Health, University of California, Los Angeles,* [e]*cbatteate@ucla.edu*

[5]*Department of Biostatistics, University of California, Los Angeles,* [f]*sudipto@ucla.edu*

The majority of Americans fail to achieve recommended levels of physical activity, which leads to numerous preventable health problems, such as diabetes, hypertension, and heart diseases. This has generated substantial interest in monitoring human activity to gear interventions toward environmental features that may relate to higher physical activity. Wearable devices, such as wrist-worn sensors that monitor gross motor activity (actigraph units) continuously record the activity levels of a subject, producing massive amounts of high-resolution measurements. Analyzing actigraph data needs to account for spatial and temporal information on trajectories or paths traversed by subjects wearing such devices. Inferential objectives include estimating a subject's physical activity levels along a given trajectory, identifying trajectories that are more likely to produce higher levels of physical activity for a given subject, and predicting expected levels of physical activity in any proposed new trajectory for a given set of health attributes. Here, we devise a Bayesian hierarchical modeling framework for spatial-temporal actigraphy data to deliver fully model-based inference on trajectories while accounting for subject-level health attributes and spatial-temporal dependencies. We undertake a comprehensive analysis of an original dataset from the Physical Activity through Sustainable Transport Approaches in Los Angeles (PASTA-LA) study to ascertain spatial zones and trajectories exhibiting significantly higher levels of physical activity while accounting for various sources of heterogeneity.

REFERENCES

ALAIMO DI LORO, P., MINGIONE, M., LIPSITT, J., BATTEATE, C. M., JERRETT, M. and BANERJEE, S. (2023). Supplement to "Bayesian hierarchical modeling and analysis for actigraph data from wearable devices." https://doi.org/10.1214/23-AOAS1742SUPPA, https://doi.org/10.1214/23-AOAS1742SUPPB

BAI, J., SUN, Y., SCHRACK, J. A., CRAINICEANU, C. M. and WANG, M.-C. (2018). A two-stage model for wearable device data. *Biometrics* **74** 744–752. MR3825361 https://doi.org/10.1111/biom.12781

BAMMANN, K., THOMSON, N. K., ALBRECHT, B. M., BUCHAN, D. S. and EASTON, C. (2021). Generation and validation of ActiGraph GT3X+ accelerometer cut-points for assessing physical activity intensity in older adults. The OUTDOOR ACTIVE validation study. *PLoS ONE* **16** e0252615. https://doi.org/10.1371/journal.pone.0252615

BANERJEE, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Anal.* **12** 583–614. MR3654826 https://doi.org/10.1214/17-BA1056R

BULL, F. C., AL-ANSARI, S. S., BIDDLE, S., BORODULIN, K., BUMAN, M. P., CARDON, G., CARTY, C., CHAPUT, J.-P., CHASTIN, S. et al. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br. J. Sports Med.* **54** 1451–1462.

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

Key words and phrases. Bayesian hierarchical models, directed acyclic graph, Gaussian processes, physical activity, sparsity, spatial-temporal statistics.

CROUTER, S. E., CLOWERS, K. G. and BASSETT JR, D. R. (2006). A novel method for using accelerometer data to predict energy expenditure. *J. Appl. Physiol.* **100** 1324–1331.

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

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. MR3553225 https://doi.org/10.1214/16-AOAS931

DE OLIVEIRA, V. (2005). Bayesian inference and prediction of Gaussian random fields based on censored data. *J. Comput. Graph. Statist.* **14** 95–115. MR2137892 https://doi.org/10.1198/106186005X27518

DOHERTY, A., JACKSON, D., HAMMERLA, N., PLÖTZ, T., OLIVIER, P., GRANAT, M. H., WHITE, T., VAN HEES, V. T., TRENELL, M. I. et al. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE* **12**.

DREWNOWSKI, A., BUSZKIEWICZ, J., AGGARWAL, A., ROSE, C., GUPTA, S. and BRADSHAW, A. (2020). Obesity and the built environment: A reappraisal. *Obesity* **28** 22–30.

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

FINLEY, A., DATTA, A. and BANERJEE, S. (2017). Spnngp: Spatial regression models for large datasets using nearest neighbor Gaussian processes. R Package Version 0.1 1.

FINLEY, A. O., DATTA, A., COOK, B. D., MORTON, D. C., ANDERSEN, H. E. and BANERJEE, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Statist.* **28** 401–414. MR3974889 https://doi.org/10.1080/10618600.2018.1537924

FREEDSON, P., BOWLES, H. R., TROIANO, R. and HASKELL, W. (2012). Assessment of physical activity using wearable monitors: Recommendations for monitor calibration and use in the field. *Med. Sci. Sports Exerc.* **44** S1–S4. https://doi.org/10.1249/MSS.0b013e3182399b7e

GELFAND, A. E., DIGGLE, P., GUTTORP, P. and FUENTES, M. (2010). *Handbook of Spatial Statistics*. CRC press.

GILKS, W. R. and ROBERTS, G. O. (1996). Strategies for improving MCMC. *Markov Chain Monte Carlo in Practice* **6** 89–114.

GOODMAN, T. and HARDIN, D. (2006). Refinable multivariate spline functions. In *Topics in Multivariate Approximation and Interpolation. Stud. Comput. Math.* **12** 55–83. Elsevier, Amsterdam. MR2410695 https://doi.org/10.1016/S1570-579X(06)80005-4

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504 https://doi.org/10.2307/3318737

HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting. *Statist. Sci.* **15** 196–223. MR1820768 https://doi.org/10.1214/ss/1009212815

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

HEDLEY, S. L. and BUCKLAND, S. T. (2004). Spatial models for line transect sampling. *J. Agric. Biol. Environ. Stat.* **9** 181–199.

JAMES, P., JANKOWSKA, M., MARX, C., HART, J. E., BERRIGAN, D., KERR, J., HURVITZ, P. M., HIPP, J. A. and LADEN, F. (2016). "Spatial energetics": Integrating data from GPS, accelerometry, and GIS to address obesity and inactivity. *Am. J. Prev. Med.* **51** 792–800. https://doi.org/10.1016/j.amepre.2016.06.006

KATZFUSS, M. and GUINNESS, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statist. Sci.* **36** 124–141. MR4194207 https://doi.org/10.1214/19-STS755

KATZFUSS, M., GUINNESS, J., GONG, W. and ZILBER, D. (2020). Vecchia approximations of Gaussian-process predictions. *J. Agric. Biol. Environ. Stat.* **25** 383–414. MR4139037 https://doi.org/10.1007/s13253-020-00401-7

KESTENS, Y., WASFI, R., NAUD, A. and CHAIX, B. (2017). "Contextualizing context": Reconciling environmental exposures, social networks, and location preferences in health research. *Curr. Environ. Health Rep.* **4** 51–60. https://doi.org/10.1007/s40572-017-0121-8

KHUSAINOV, R., AZZI, D., ACHUMBA, I. E. and BERSCH, S. D. (2013). Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations. *Sensors* **13** 12852–12902. https://doi.org/10.3390/s131012852

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

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

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653 https://doi.org/10.1093/biomet/81.1.27

LYDEN, K., KEADLE, S. K., STAUDENMAYER, J. and FREEDSON, P. S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Med. Sci. Sports Exerc.* **46** 386–397. https://doi.org/10.1249/MSS.0b013e3182a42a2d

MADDISON, R., HOORN, S. V., JIANG, Y., MHURCHU, C. N., EXETER, D., DOREY, E., BULLEN, C., UTTER, J., SCHAAF, D. et al. (2009). The environment and physical activity: The influence of psychosocial, perceived and built environmental factors. *Int. J. Behav. Nutr. Phys. Act.* **6** 19.

MATHIE, M. J., COSTER, A. C. F., LOVELL, N. H. and CELLER, B. G. (2003). Detection of daily physical activities using a triaxial accelerometer. *Med. Biol. Eng. Comput.* **41** 296–301. https://doi.org/10.1007/BF02348434

MIGUELES, J. H., CADENAS-SANCHEZ, C., EKELUND, U., NYSTRÖM, C. D., MORA-GONZALEZ, J., LÖF, M., LABAYEN, I., RUIZ, J. R. and ORTEGA, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Med.* **47** 1821–1845. https://doi.org/10.1007/s40279-017-0716-0

MIGUELES, J. H., CADENAS-SANCHEZ, C., ROWLANDS, A. V., HENRIKSSON, P., SHIROMA, E. J., ACOSTA, F. M., RODRIGUEZ-AYLLON, M., ESTEBAN-CORNEJO, I., PLAZA-FLORIDO, A. et al. (2019). Comparability of accelerometer signal aggregation metrics across placements and dominant wrist cut points for the assessment of physical activity in adults. *Sci. Rep.* **9** 1–12.

MOLSTAD, A. J., HSU, L. and SUN, W. (2021). Gaussian process regression for survival time prediction with genome-wide gene expression. *Biostatistics* **22** 164–180. MR4207149 https://doi.org/10.1093/biostatistics/kxz023

MURPHY, K. P. (2012). *Machine Learning*: *A Probabilistic Perspective*. MIT Press, Cambridge.

PERUZZI, M., BANERJEE, S. and FINLEY, A. O. (2022). Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *J. Amer. Statist. Assoc.* **117** 969–982. MR4436326 https://doi.org/10.1080/01621459.2020.1833889

PIERCY, K. L., TROIANO, R. P., BALLARD, R. M., CARLSON, S. A., FULTON, J. E., GALUSKA, D. A., GEORGE, S. M. and OLSON, R. D. (2018). The physical activity guidelines for Americans. *JAMA* **320** 2020–2028.

PLASQUI, G. and WESTERTERP, K. R. (2007). Physical activity assessment with accelerometers: An evaluation against doubly labeled water. *Obesity* **15** 2371–2379. https://doi.org/10.1038/oby.2007.281

RAMSAY, J. O. and SILVERMAN, B. W. (2007). *Applied Functional Data Analysis*: *Methods and Case Studies*. Springer, Berlin.

RAY, E. L., SASAKI, J. E., FREEDSON, P. S. and STAUDENMAYER, J. (2018). Physical activity classification with dynamic discriminative methods. *Biometrics* **74** 1502–1511. MR3908166 https://doi.org/10.1111/biom.12892

REINER, M., NIERMANN, C., JEKAUC, D. and WOLL, A. (2013). Long-term health benefits of physical activity–a systematic review of longitudinal studies. *BMC Public Health* **13** 1–9.

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

SIKKA, R. S., BAER, M., RAJA, A., STUART, M. and TOMPKINS, M. (2019). Analytics in sports medicine: Implications and responsibilities that accompany the era of big data. *JBJS* **101** 276–283.

STAUDENMAYER, J., HE, S., HICKEY, A., SASAKI, J. and FREEDSON, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J. Appl. Physiol.* **119** 396–403.

STEIN, M. L., CHI, Z. and WELTY, L. J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 275–296. MR2062376 https://doi.org/10.1046/j.1369-7412.2003.05512.x

TARALDSEN, K., CHASTIN, S. F., RIPHAGEN, I. I., VEREIJKEN, B. and HELBOSTAD, J. L. (2012). Physical activity monitoring by use of accelerometer-based body-worn sensors in older adults: A systematic literature review of current knowledge and applications. *Maturitas* **71** 13–19.

VAN HEES, V. T., RENSTRÖM, F., WRIGHT, A., GRADMARK, A., CATT, M., CHEN, K. Y., LÖF, M., BLUCK, L., POMEROY, J. et al. (2011). Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. *PLoS ONE* **6** e22922.

VAN LOO, C. M., OKELY, A. D., BATTERHAM, M. J., HINKLEY, T., EKELUND, U., BRAGE, S., REILLY, J. J., TROST, S. G., JONES, R. A. et al. (2018). Wrist acceleration cut-points for moderate-to-vigorous physical activity in youth. *Med. Sci. Sports Exerc.* **50** 609.

VECCHIA, A. V. (1988). Estimation and model identification for continuous spatial processes. *J. Roy. Statist. Soc. Ser. B* **50** 297–312. MR0964183

WHITE, T., WESTGATE, K., WAREHAM, N. J. and BRAGE, S. (2016). Estimation of physical activity energy expenditure during free-living from wrist accelerometry in UK adults. *PLoS ONE* **11** e0167472. https://doi.org/10.1371/journal.pone.0167472

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2887–2902
https://doi.org/10.1214/23-AOAS1743

BAYESIAN LEARNING OF COVID-19 VACCINE SAFETY WHILE INCORPORATING ADVERSE EVENTS ONTOLOGY

BY BANGYAO ZHAO[a], YUAN ZHONG[b], JIAN KANG[c] AND LILI ZHAO[d]

Department of Biostatistics, University of Michigan, [a]*byzhao@umich.edu,* [b]*ylzhong@umich.edu,* [c]*jiankang@umich.edu,* [d]*zhaolili@med.umich.edu*

While vaccines are crucial to end the COVID-19 pandemic, public confidence in vaccine safety has always been vulnerable. Many statistical methods have been applied to VAERS (Vaccine Adverse Event Reporting System) database to study the safety of COVID-19 vaccines. However, none of these methods considered the adverse event (AE) ontology. AEs are naturally related; for example, events of retching, dysphagia, and reflux are all related to an abnormal digestive system. Explicitly bringing AE relationships into the model can aid in the detection of true AE signals amid the noise while reducing false positives. We propose a Bayesian graph-assisted signal selection (BGrass) model to simultaneously estimate all AEs while incorporating the network of dependence between AEs. Under a fully Bayesian inference framework, we also propose a negative control approach to mitigate the reporting bias and an enrichment approach to detecting AE groups of concern. For posterior computation we construct an equivalent model representation and develop an efficient Gibbs sampler. We evaluate the performance of BGrass via extensive simulations. To study the safety of COVID-19 vaccines, we apply BGrass to analyze approximately one million VAERS reports (01/01/2016–12/24/2021) involving more than 800 AEs. In particular, we found that blood clots (including deep vein thrombosis, thrombosis, and pulmonary embolism) are more likely to be reported after COVID-19 vaccination, compared to influenza vaccines. They are also reported more often for Johnson & Johnson–Janssen vaccine, compared to mRNA-based COVID-19 vaccines. A user-friendly `R` package `BGrass` that implements the proposed methods to assess vaccine safety is included in the Supplementary Material and is publicly available at https://github.com/BangyaoZhao/BGrass.

REFERENCES

BATE, A., LINDQUIST, M., EDWARDS, I. R., OLSSON, S., ORRE, R., LANSNER, A. and FREITAS, R. M. D. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *Eur. J. Clin. Pharmacol.* **54** 315–321.

BIYKEM, B., ISHAN, K. and PETER, J. H. (2021). Myocarditis with Covid-19 mRNA vaccines. *Circulation* **144** 471–484.

CHUNG, F. R. K. (1997). *Spectral Graph Theory*. *CBMS Regional Conference Series in Mathematics* **92**. Amer. Math. Soc., Providence, RI. MR1421568

DUMOUCHEL, W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Amer. Statist.* **53** 177–190.

DUMOUCHEL, W. and PREGIBON, D. (2001). Empirical Bayes screening for multi-item associations. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 67–76. San Francisco,CA.

EDELMAN, A., BONIFACE, E. R., BENHAR, E., HAN, L., MATTESON, K. A., FAVARO, C., PEARSON, J. T. and DARNEY, B. G. (2022). Association between menstrual cycle length and coronavirus disease 2019 (Covid-19) vaccination. *Obstet. Gynecol.*.

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

Key words and phrases. Bayesian model, adverse event ontology, COVID-19 vaccine, graphical model, negative control.

HIPPISLEY-COX, J., PATONE, M., MEI, X. W., SAATCI, D., DIXON, S., KHUNTI, K., ZACCARDI, F., WATKINSON, P., SHANKAR-HARI, M. et al. (2021). Risk of thrombocytopenia and thromboembolism after Covid-19 vaccination and Sars-CoV-2 positive testing: Self-controlled case series study. *BMJ* **374** n1931. https://doi.org/10.1136/bmj.n1931

HUANG, L., ZALKIKAR, J. and TIWARI, R. C. (2011). A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *J. Amer. Statist. Assoc.* **106** 1230–1241. MR2896832 https://doi.org/10.1198/jasa.2011.ap10243

HUANG, L., ZALKIKAR, J. and TIWARI, R. (2014). Likelihood ratio based tests for longitudinal drug safety data. *Stat. Med.* **33** 2408–2424. MR3256675 https://doi.org/10.1002/sim.6103

HUANG, L., GUO, T., ZALKIKAR, J. N. and TIWARI, R. C. (2014). A review of statistical methods for safety surveillance. *Ther. Innov. Regul. Sci.* **48** 98–108.

JABAGI, M. J., BOTTON, J., BERTRAND, M., WEILL, A., FARRINGTON, P., ZUREIK, M. and DRAY-SPIRA, R. (2022). Myocardial infarction, stroke, and pulmonary embolism after BNT162b2 mRNA Covid-19 vaccine in people aged 75 years or older. *JAMA* **327** 80–82.

KADALI, R. A. K., JANAGAMA, R., PERURU, S. and MALAYALA, S. V. (2021). Side effects of BNT162b2 mRNA Covid-19 vaccine: A randomized, cross-sectional study with detailed self-reported symptoms from healthcare workers. *Int. J. Infect. Dis.* **106** 376–381. https://doi.org/10.1016/j.ijid.2021.04.047

LANSNER, R. O. A., BATE, A., LINDQUIST, I. R. E. M., OLSSON, S., ORRE, R., LANSNER, A. and FREITAS, R. M. D. (2000). Bayesian neural networks with confidence estimations applied to data mining. *Comput. Statist. Data Anal.* **34** 473–493.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LI, S. and ZHAO, L. (2021). Vaccine adverse event enrichment tests. *Stat. Med.* **40** 4269–4278. MR4300085 https://doi.org/10.1002/sim.9027

MALEKI, F., OVENS, K., HOGAN, D. J. and KUSALIK, A. J. (2020). Gene set analysis: Challenges, opportunities, and future research. *Front. Genet.* **11** 654. https://doi.org/10.3389/fgene.2020.00654

MOORE, N., HALL, G., STURKENBOOM, M., MANN, R., LAGNAOUI, R. and BEGAUD, B. (2003). Biases affecting the proportional reporting ratio (PRR) in spontaneous reports pharmacovigilance databases: The example of sertindole. *Pharmacoepidemiol. Drug Saf.* **12** 271–281.

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

NORÉN, G. N., BATE, A., ORRE, R. and EDWARDS, I. R. (2006). Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat. Med.* **25** 3740–3757. MR2252427 https://doi.org/10.1002/sim.2473

OSTER, M. E., SHAY, D. K., SU, J. R., GEE, J., CREECH, C. B., BRODER, K. R., EDWARDS, K., SOSLOW, J. H., DENDY, J. M. et al. (2022). Myocarditis cases reported after mRNA-based Covid-19 vaccination in the US from December 2020 to August 2021. *JAMA* **327** 331–340.

PAPAMANOLI, A., THORNE, M. and PSEVDOS, G. (2021). Delayed skin rash after receiving Sars-CoV-2 mRNA moderna vaccine. *Infect. Dis. Clin. Pract.* **29** 262–263.

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

RAMALINGAM, S., ARORA, H., LEWIS, S., GUNASEKARAN, K., MURUGANANDAM, M., NAGARAJU, S. and PADMANABHAN, P. (2021). Covid-19 vaccine-induced cellulitis and myositis. *Clev. Clin. J. Med.* **88** 648–650. https://doi.org/10.3949/ccjm.88a.21038

ROTHMAN, K. J., LANES, S. and SACKS, S. T. (2004). The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol. Drug Saf.* **13** 519–523. https://doi.org/10.1002/pds.1001

SCHUEMIE, M. J., RYAN, P. B., DUMOUCHEL, W., SUCHARD, M. A. and MADIGAN, D. (2014). Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Stat. Med.* **33** 209–218. MR3146759 https://doi.org/10.1002/sim.5925

SEE, I., SU, J. R., LALE, A., WOO, E. J., GUH, A. Y., SHIMABUKURO, T. T., STREIFF, M. B., RAO, A. K., WHEELER, A. P. et al. (2021). US case reports of cerebral venous sinus thrombosis with thrombocytopenia after Ad26. COV2. S vaccination, March 2 to April 21, 2021. *JAMA*.

SHI, X., MIAO, W. and TCHETGEN, E. T. (2020). A selective review of negative control methods in epidemiology. *Curr. Epidemiol. Rep.* **7** 190–202. https://doi.org/10.1007/s40471-020-00243-4

SHIMABUKURO, T. T., COLE, M. and SU, J. R. (2021). Reports of anaphylaxis after receipt of mRNA Covid-19 vaccines in the US-December 14, 2020-January 18, 2021. *JAMA* **325** 1101–1102. https://doi.org/10.1001/jama.2021.1967

SHIMABUKURO, T. T., NGUYEN, M., MARTIN, D. and DESTEFANO, F. (2015). Safety monitoring in the vaccine adverse event reporting system (VAERS). *Vaccine* **33** 4398–4405. https://doi.org/10.1016/j.vaccine.2015.07.035

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

SUN, H. and LI, H. (2010). A Bayesian approach for graph-constrained estimation for high-dimensional regression. *Int. J. Syst. Synth. Biol.* **1** 255.

WOO, E. J., MBA-JONAS, A., DIMOVA, R. B., ALIMCHANDANI, M., ZINDERMAN, C. E. and NAIR, N. (2021). Association of receipt of the Ad26.COV2.S Covid-19 vaccine with presumptive Guillain–Barré syndrome, February–July 2021. *JAMA* **326** 1606–1613. https://doi.org/10.1001/jama.2021.16496

ZHAO, L., LEE, S., LI, R., ONG, E., HE, Y. and FREED, G. (2020). Improvement in the analysis of vaccine adverse event reporting system database. *Stat. Biopharm. Res.* **12** 303–310.

ZHAO, B., ZHONG, Y., KANG, J. and ZHAO, L. (2023). Supplement to "Bayesian learning of COVID-19 Vaccine safety while incorporating adverse events ontology." https://doi.org/10.1214/23-AOAS1743SUPPA, https://doi.org/10.1214/23-AOAS1743SUPPB, https://doi.org/10.1214/23-AOAS1743SUPPC, https://doi.org/10.1214/23-AOAS1743SUPPD

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2903–2923
https://doi.org/10.1214/23-AOAS1744

ADDRESSING SELECTION BIAS AND MEASUREMENT ERROR IN COVID-19 CASE COUNT DATA USING AUXILIARY INFORMATION

BY WALTER DEMPSEY[a]

Department of Biostatistics, University of Michigan, [a]*wdem@umich.edu*

Coronavirus case-count data has influenced government policies and drives most epidemiological forecasts. Limited testing is cited as the key driver behind minimal information on the COVID-19 pandemic. While expanded testing is laudable, measurement error and selection bias are the two greatest problems limiting our understanding of the COVID-19 pandemic; neither can be fully addressed by increased testing capacity. In this paper we demonstrate their impact on estimation of point prevalence and the effective reproduction number. We show that estimates, based on the millions of molecular tests in the U.S., have the same mean square error as a small simple random sample. To address this, a procedure is presented that combines case-count data and random samples over time to estimate selection propensities based on key covariate information. We then combine these selection propensities with epidemiological forecast models to construct a *doubly robust* estimation method that accounts for both measurement-error and selection bias. This method is then applied to estimate Indiana's active infection prevalence using case-count, hospitalization, and death data with demographic information, a statewide random molecular sample collected from April 25–29, 2020, and Delphi's COVID-19 Trends and Impact Survey. We end with a series of recommendations based on the proposed methodology.

REFERENCES

ACCORSI, E. K., QIU, X., RUMPLER, E., KENNEDY-SHAFFER, L., KAHN, R., JOSHI, K., GOLDSTEIN, E., STENSRUD, M. J., NIEHUS, R. et al. (2021). How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. *Eur. J. Epidemiol*. **36** 179–196. https://doi.org/10.1007/s10654-021-00727-7

ADAMS, D. (2020). Coronavirus testing in Indiana: Here's who can get a test. https://www.indystar.com/story/news/health/2020/05/12/coronavirus-testing-indiana-who-should-get-tested/3110592001/.

AREVALO-RODRIGUEZ, I., BUITRAGO-GARCIA, D., SIMANCAS-RACINES, D., ZAMBRANO-ACHIG, P., DEL CAMPO, R., CIAPPONI, A., SUED, O., MARTINEZ-GARCÍA, L., RUTJES, A. W. et al. (2020). False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS ONE* **15** 1–19.

BARKAY, N., COBB, C., EILAT, R., GALILI, T., HAIMOVICH, D., LAROCCA, S., MORRIS, K. and SARIG, T. (2020). Weights and Methodology Brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in Partnership with Facebook.

BEESLEY, L. J., FRITSCHE, L. G. and MUKHERJEE, B. (2020). An analytic framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *Stat. Med.* **39** 1965–1979. MR4105271 https://doi.org/10.1002/sim.8524

BEESLEY, L. J. and MUKHERJEE, B. (2019). Statistical inference for association studies using electronic health records: Handling both selection bias and outcome misclassification. *MedRxiv*.

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

CHEN, Y., LI, P. and WU, C. (2020). Doubly robust inference with nonprobability survey samples. *J. Amer. Statist. Assoc.* **115** 2011–2021. MR4189773 https://doi.org/10.1080/01621459.2019.1677241

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0474575

COHEN, A. N., KESSEL, B. and MILGROOM, M. G. (2020). Diagnosing COVID-19 infection: The danger of over-reliance on positive test results. *MedRxiv*.

Key words and phrases. Selection bias, measurement error, inverse-probability weighting, doubly robust estimation, COVID-19.

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. https://doi.org/10.1093/aje/kwq084

CORI, A., FERGUSON, N. M., FRASER, C. and CAUCHEMEZ, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *Amer. J. Epidemiol.* **178** 1505–1512. https://doi.org/10.1093/aje/kwt133

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. CRC Press, London. MR0370837

DEMPSEY, W. (2023). Supplement to "Addressing selection bias and measurement error in COVID-19 case count data using auxiliary information." https://doi.org/10.1214/23-AOAS1744SUPPA, https://doi.org/10.1214/23-AOAS1744SUPPB

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

ELLIOTT, M. R. and VALLIANT, R. (2017). Inference for nonprobability samples. *Statist. Sci.* **32** 249–264. MR3648958 https://doi.org/10.1214/16-STS598

FOX, M. P., LASH, T. L. and BODNAR, L. M. (2020). Common misconceptions about validation studies. *Int. J. Epidemiol.* **49** 1392–1396. https://doi.org/10.1093/ije/dyaa090

FRASER, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE* **2** e758.

GIORDANO, G., BLANCHINI, F., BRUNO, R., COLANERI, P., FILIPPO, A. D., MATTEO, A. D. and COLANERI, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nat. Med.* **26** 855–860. https://doi.org/10.1038/s41591-020-0883-7

HAO, X., CHENG, S., WU, D., WU, T., LIN, X. and WANG, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* 420–424.

HENG, K. and ALTHAUS, C. L. (2020). The approximately universal shapes of epidemic curves in the Susceptible-Exposed-Infectious-Recovered (SEIR) model. *Sci. Rep.* **10** 19365. https://doi.org/10.1038/s41598-020-76563-8

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

IDOH (2021). COVID-19 Case Demographics Daily Trend. Available at https://hub.mph.in.gov/dataset/covid-19-case-demographics-daily-trend/resource/c8a0ff06-7ff6-4932-b61e-a87ad2710797. Accessed: 2021-06-15.

IHME and MURRAY, C. J. (2020). Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European Economic Area countries. *MedRxiv*.

IRONS, N. J. and RAFTERY, A. E. (2021). Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc. Natl. Acad. Sci. USA* **118**.

JOHNDROW, J., LUM, K., GARGIULO, M. and BALL, P. (2020). Estimating the number of SARS-CoV-2 infections and the impact of social distancing in the United States.

KAHN, R., KENNEDY-SHAFFER, L., GRAD, Y., ROBINS, J. and LIPSITCH, M. (2021). Potential biases arising from epidemic dynamics in observational seroprotection studies. *Amer. J. Epidemiol.* **192** 328–335.

KATZ, A., CIVANTOS, F., SARGI, Z., LEIBOWITZ, J., NICOLLI, E., WEED, D., MOSKOVITZ, A., CIVANTOS, A., ANDREWS, D. et al. (2020). False-positive reverse transcriptase polymerase chain reaction screening for SARS-CoV-2 in the setting of urgent head and neck surgery and otolaryngologic emergencies during the pandemic: Clinical implications. *Head Neck* **42** 1621–1628.

KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587 https://doi.org/10.1111/rssa.12136

LAUER, S., GRANTZ, K., QIFANG, B., JONES, F., ZHENG, Q., MEREDITH, H., AZMAN, A., REICH, N. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172** 577–582.

LEUNG, G. (2020). Lockdown can't last forever. Here's how to lift it. *N.Y. Times*.

LEVIN, A. T., HANAGE, W. P., OWUSU-BOAITEY, N., COCHRAN, K. B., WALSH, S. P. and MEYEROWITZ-KATZ, G. (2020). Assessing the age specificity of infection fatality rates for Covid-19: Systematic review, meta-analysis, and public policy implications. *MedRxiv*.

MAY, E. (2020). Each of Indiana's reopening stages, explained. https://www.indystar.com/story/news/health/2020/05/01/when-indiana-reopen-here-phases-set-reopening/3067992001/.

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat.* **12** 685–726. MR3834282 https://doi.org/10.1214/18-AOAS1161SF

MINA, M. J., PARKER, R. and LARREMORE, D. B. (2020). Rethinking Covid-19 test sensitivity—a strategy for containment. *N. Engl. J. Med.* **383** e120. https://doi.org/10.1056/NEJMp2025631

NEWMAN, M. E. J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E* (3) **66** 016128. MR1919737 https://doi.org/10.1103/PhysRevE.66.016128

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

PARSHANI, R., CARMI, S. and HAVLIN, S. (2010). Epidemic threshold for the susceptible-infectious-susceptible model on random networks. *Phys. Rev. Lett.* **104** 258701. https://doi.org/10.1103/PhysRevLett.104.258701

PASTOR-SATORRAS, R. and VESPIGNANI, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86** 3200–3203. https://doi.org/10.1103/PhysRevLett.86.3200

RAY, D., SALVATORE, M., BHATTACHARYYA, R., WANG, L., MOHAMMED, S., PURKAYASTHA, S., HALDER, A., RIX, A., BARKER, D. et al. (2020). Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: Data science call to arms. *MedRxiv*.

REPORTS, S. (2020). Indiana opens up COVID-19 testing to all Hoosiers with symptoms. https://www.wishtv.com/news/medical/indiana-opens-up-covid-19-testing-to-more-hoosiers/.

RUDAVSKY, S. (2020). Want a coronavirus test? Anyone can get one now, state says. Here's how. https://www.indystar.com/story/news/health/2020/06/12/indiana-says-anyone-who-wants-coronavirus-test-can-get-one/3179151001/.

SALOMON, J. A., REINHART, A., BILINSKI, A., CHUA, E. J., MOTTE-KERR, W. L., RÖNN, M. M., REITSMA, M. B., MORRIS, K. A., LAROCCA, S. et al. (2021). The US COVID-19 trends and impact survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proc. Natl. Acad. Sci. USA* **118** e2111454118.

SCARPETTA, S., PEARSON, M., COLOMBO, F., GUANAI, F., DEDET, G., LOPERT, R. and WENZ, M. (2021). OECD Policy Responses to Coronavirus (COVID-19). Available at https://www.oecd.org/coronavirus/policy-responses/testing-for-covid-19-how-to-best-use-the-various-tests-c76df201/. Accessed: 2021-06-15.

SIDDARTH, D. and WEYL, E. (2020). Why we must test millions a day. *COVID-19 Rapid Response Impact Initiative*.

SMITH, M., YOURISH, K., ALMUKHTAR, S., COLLINS, K., IVORY, D. and HARMON, A. (2020). Coronavirus in the US. *N.Y. Times*.

SONG, P. X., WANG, L., ZHOU, Y., HE, J., ZHU, B., WANG, F., TANG, L. and EISENBERG, M. (2020). An epidemiological forecast model and software assessing interventions on COVID-19 epidemic in China. *MedRxiv*.

VALLIANT, R. and DEVER, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.* **40** 105–137. MR2758301 https://doi.org/10.1177/0049124110392533

VAN SMEDEN, M., LASH, T. L. and GROENWOLD, R. H. H. (2019). Reflection on modern methods: Five myths about measurement error in epidemiological research. *Int. J. Epidemiol.* **49** 338–347.

WALLINGA, J. and TEUNIS, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Amer. J. Epidemiol.* **160** 509–516. https://doi.org/10.1093/aje/kwh255

WANG, C., LIU, L., HAO, X., GUO, H., WANG, Q., HUANG, J., HE, N., YU, H., LIN, X. et al. (2020b). Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in Wuhan, China. *MedRxiv*.

WESTREICH, D., EDWARDS, J. K., LESKO, C. R., STUART, E. and COLE, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *Amer. J. Epidemiol.* **186** 1010–1014. https://doi.org/10.1093/aje/kwx164

WESTREICH, D., EDWARDS, J., LESKO, C., COLE, S. and STUART, E. (2018). Target validity and the hierarchy of study designs. *Amer. J. Epidemiol.* **188**.

WOLOSHIN, S., PATEL, N. and KESSELHEIM, A. S. (2020). False negative tests for SARS-CoV-2 infection–challenges and implications. *N. Engl. J. Med.* **383** e38. https://doi.org/10.1056/NEJMp2015897

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

YIANNOUTSOS, C. T., HALVERSON, P. K. and MENACHEMI, N. (2021). Bayesian estimation of SARS-CoV-2 prevalence in Indiana by random testing. *Proc. Natl. Acad. Sci. USA* **118** e2013906118. MR4275060 https://doi.org/10.1073/pnas.2013906118

ZHAO, Q., JU, N., BACALLADO, S. and SHAH, R. D. (2021). BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *Ann. Appl. Stat.* **15** 363–390. MR4255283 https://doi.org/10.1214/20-aoas1401

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2924–2943
https://doi.org/10.1214/23-AOAS1745

PAIRWISE NONLINEAR DEPENDENCE ANALYSIS OF GENOMIC DATA

BY SIQI XIANG[1,a], WAN ZHANG[1,b], SIYAO LIU[2,e], KATHERINE A. HOADLEY[2,f], CHARLES M. PEROU[2,g], KAI ZHANG[1,c] AND J. S. MARRON[1,d]

[1]*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill,* [a]*xiangsiqi.unc@gmail.com,* [b]*wan.zhang@unc.edu,* [c]*zhangk@email.unc.edu,* [d]*marron@unc.edu*

[2]*Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill,* [e]*siyao@email.unc.edu,* [f]*hoadley@med.unc.edu,* [g]*cperou@med.unc.edu*

In The Cancer Genome Atlas (TCGA) data set, there are many interesting nonlinear dependencies between pairs of genes that reveal important relationships and subtypes of cancer. Such genomic data analysis requires a rapid, powerful, and interpretable detection process, especially in a high-dimensional environment. We study the nonlinear patterns among the expression of pairs of genes from TCGA using a powerful tool called binary expansion testing. We find many nonlinear patterns, some of which are driven by known cancer subtypes, some of which are novel.

REFERENCES

BRUEFFER, C., VALLON-CHRISTERSSON, J., GRABAU, D., EHINGER, A., HÄKKINEN, J., HEGARDT, C., MALINA, J., CHEN, Y., BENDAHL, P.-O. et al. (2018). Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: A report from the population-based multicenter Sweden cancerome analysis network—breast initiative. *JCO Precis. Oncol.* **2** 1–18. https://doi.org/10.1200/PO.17.00135

CHARAFE-JAUFFRET, E., GINESTIER, C., MONVILLE, F., FINETTI, P., ADELAÏDE, J., CERVERA, N., FEKAIRI, S., XERRI, L., JACQUEMIER, J. et al. (2006). Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene* **25** 2273–2284. https://doi.org/10.1038/sj.onc.1209254

CIRIELLO, G., GATZA, M. L., BECK, A. H., WILKERSON, M. D., RHIE, S. K., PASTORE, A., ZHANG, H., MCLELLAN, M., YAU, C. et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163** 506–519. https://doi.org/10.1016/j.cell.2015.09.033

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. MR0029139 https://doi.org/10.1214/aoms/1177730150

IGLESIA, M. D., PARKER, J. S., HOADLEY, K. A., SERODY, J. S., PEROU, C. M. and VINCENT, B. G. (2016). Genomic analysis of immune cell infiltrates across 11 tumor types. *J. Natl. Cancer Inst.* **108**. https://doi.org/10.1093/jnci/djw144

KAC, M. (1959). *Statistical Independence in Probability, Analysis and Number Theory*. *The Carus Mathematical Monographs* **12**. Math. Assoc. of America; distributed by Wiley, New York. MR0110114

KINNEY, J. B. and ATWAL, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **111** 3354–3359. MR3200177 https://doi.org/10.1073/pnas.1309933111

KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* (3) **69** 066138, 16. MR2096503 https://doi.org/10.1103/PhysRevE.69.066138

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDÓTTIR, H., TAMAYO, P. and MESIROV, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27** 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

LIBERZON, A., BIRGER, C., THORVALDSDÓTTIR, H., GHANDI, M., MESIROV, J. P. and TAMAYO, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1** 417–425. https://doi.org/10.1016/j.cels.2015.12.004

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

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

Key words and phrases. Binary expansion, genomic data, nonlinear dependence, nonparametric dependence testing.

SAAL, L., VALLON-CHRISTERSSON, J., HÄKKINEN, J., HEGARDT, C., GRABAU, D., WINTER, C., BRUEFFER, C., TANG, M.-H. E., REUTERSWÄRD, C. et al. (2015). The Sweden Cancerome Analysis Network—Breast (SCAN-B) initiative: A large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. *Gen. Med.* **7** 20. https://doi.org/10.1186/s13073-015-0131-9

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550. https://doi.org/10.1073/pnas.0506580102

SUN, N. and ZHAO, H. (2014). Putting things in order. *Proc. Natl. Acad. Sci. USA* **111** 16236–16237. https://doi.org/10.1073/pnas.1418862111

SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745 https://doi.org/10.1016/j.jspi.2013.03.018

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

THE CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

WILKINSON, L., ANAND, A. and GROSSMAN, R. (2005). Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization* 157–164. IEEE Comput. Soc., Los Alamitos.

XIANG, S., ZHANG, W., LIU, S., HOADLEY, K. A., PEROU, C. M., ZHANG, K. and MARRON, J. S. (2023). Supplement to "Pairwise nonlinear dependence analysis of genomic data." https://doi.org/10.1214/23-AOAS1745SUPPA, https://doi.org/10.1214/23-AOAS1745SUPPB, https://doi.org/10.1214/23-AOAS1745SUPPC, https://doi.org/10.1214/23-AOAS1745SUPPD, https://doi.org/10.1214/23-AOAS1745SUPPE

ZHANG, K. (2019). BET on independence. *J. Amer. Statist. Assoc.* **114** 1620–1637. MR4047288 https://doi.org/10.1080/01621459.2018.1537921

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2944–2969
https://doi.org/10.1214/23-AOAS1746

GENERALIZED MATRIX DECOMPOSITION REGRESSION: ESTIMATION AND INFERENCE FOR TWO-WAY STRUCTURED DATA

BY YUE WANG[1,a], ALI SHOJAIE[2,b], TIMOTHY RANDOLPH[3,c], PARKER KNIGHT[4,d] AND JING MA[5,e]

[1]*Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus,* [a]*yue.2.wang@cuanschutz.edu*

[2]*Department of Biostatistics, University of Washington,* [b]*ashojaie@uw.edu*

[3]*Clinical Research Division, Fred Hutchinson Cancer Center,* [c]*trandolp@fredhutch.org*

[4]*Department of Biostatistics, Harvard University,* [d]*pknight@g.harvard.edu*

[5]*Public Health Sciences Division, Fred Hutchinson Cancer Center,* [e]*jingma@fredhutch.org*

Motivated by emerging applications in ecology, microbiology, and neuroscience, this paper studies high-dimensional regression with two-way structured data. To estimate the high-dimensional coefficient vector, we propose the generalized matrix decomposition regression (GMDR) to efficiently leverage auxiliary information on row and column structures. GMDR extends the principal component regression (PCR) to two-way structured data, but unlike PCR, GMDR selects the components that are most predictive of the outcome, leading to more accurate prediction. For inference on regression coefficients of individual variables, we propose the generalized matrix decomposition inference (GMDI), a general high-dimensional inferential framework for a large family of estimators that include the proposed GMDR estimator. GMDI provides more flexibility for incorporating relevant auxiliary row and column structures. As a result, GMDI does not require the true regression coefficients to be sparse but constrains the coordinate system representing the regression coefficients according to the column structure. GMDI also allows dependent and heteroscedastic observations. We study the theoretical properties of GMDI in terms of both the type-I error rate and power and demonstrate the effectiveness of GMDR and GMDI in simulation studies and an application to human microbiome data.

REFERENCES

ALLEN, G. I., GROSENICK, L. and TAYLOR, J. (2014). A generalized least-square matrix decomposition. *J. Amer. Statist. Assoc.* **109** 145–159. MR3180553 https://doi.org/10.1080/01621459.2013.852978

BÄCKHED, F., ROSWALL, J., PENG, Y., FENG, Q., JIA, H., KOVATCHEVA-DATCHARY, P., LI, Y., XIA, Y., XIE, H. et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17** 690–703.

BANA, B. and CABREIRO, F. (2019). The microbiome and aging. *Annu. Rev. Genet.* **53** 239–261. https://doi.org/10.1146/annurev-genet-112618-043650

BELKAID, Y. and HAND, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell* **157** 121–141. https://doi.org/10.1016/j.cell.2014.03.011

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097 https://doi.org/10.1093/biomet/asu056

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

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7** 335–336.

Key words and phrases. Dimensionality reduction, high-dimensional inference, microbiome data, prediction, two-way structured data.

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655 https://doi.org/10.1214/088342306000000682

CUESTA, S. M., RAHMAN, S. A., FURNHAM, N. and THORNTON, J. M. (2015). The classification and evolution of enzyme function. *Biophys. J.* **109** 1082–1086.

DOMINGUEZ-BELLO, M. G., COSTELLO, E. K., CONTRERAS, M., MAGRIS, M., HIDALGO, G., FIERER, N. and KNIGHT, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. USA* **107** 11971–11975.

ESCOUFIER, Y. (1987). The duality diagram: A means for better practical applications. In *Developments in Numerical Ecology* (*Roscoff*, 1986). *NATO Adv. Sci. Inst. Ser. G: Ecolog. Sci.* **14** 139–156. Springer, Berlin. MR0913539 https://doi.org/10.1007/978-3-642-70880-0_3

ESCOUFIER, Y. (2006). Operator related to a data matrix: A survey. In *COMPSTAT* 2006—*Proceedings in Computational Statistics* (A. Rizzi and M. Vichi, eds.) 285–297. Physica, Heidelberg. MR2330545 https://doi.org/10.1007/978-3-7908-1709-6_22

FANG, P., KAZMI, S., JAMESON, K. and HSIAO, E. (2020). The microbiome as a modifier of neurodegenerative disease risk. *Cell Host Microbe* **28** 201–222.

GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*, 4th ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR3024913

GUPTA, A. K. and NAGAR, D. K. (2018). *Matrix Variate Distributions*. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics* **104**. CRC Press/CRC, Boca Raton, FL. MR1738933

GURUNG, M., LI, Z., YOU, H., RODRIGUES, R., JUMP, D. B., MORGUN, A. and SHULZHENKO, N. (2020). Role of gut microbiota in type 2 diabetes pathophysiology. *eBioMedicine* **51** 102590.

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

HULLAR, M. A., JENKINS, I. C., RANDOLPH, T. W., CURTIS, K. R., MONROE, K. R., ERNST, T., SHEPHERD, J. A., STRAM, D. O., CHENG, I. et al. (2021). Associations of the gut microbiome with hepatic adiposity in the Multiethnic Cohort Adiposity Phenotype Study. *Gut Microbes* **13** 1965463.

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

JAVANMARD, A. and MONTANARI, A. (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629

KANEHISA, M. (2000). *Post-Genome Informatics*. Oxford Univ. Press, Oxford.

KARAS, M., BRZYSKI, D., DZEMIDZIC, M., GOÑI, J., KAREKEN, D. A., RANDOLPH, T. W. and HAREZLAK, J. (2019). Brain connectivity-informed regularization methods for regression. *Stat. Biosci.* **11** 47–90. https://doi.org/10.1007/s12561-017-9208-x

KELLY, T. N., BAZZANO, L. A., AJAMI, N. J., HE, H., ZHAO, J., PETROSINO, J. F., CORREA, A. and HE, J. (2016). Gut microbiome associates with lifetime cardiovascular disease risk profile among bogalusa heart study participants. *Circ. Res.* **119** 956–964.

LI, S., CAI, T. T. and LI, H. (2022). Inference for high-dimensional linear mixed-effects models: A quasi-likelihood approach. *J. Amer. Statist. Assoc.* **117** 1835–1846. MR4528474 https://doi.org/10.1080/01621459.2021.1888740

LI, Y., YANG, M. and ZHANG, Z. (2018). A survey of multi-view representation learning. *IEEE Trans. Knowl. Data Eng.* **31** 1863–1883.

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088. MR2414585 https://doi.org/10.1111/j.1541-0420.2007.00799.x

LOZUPONE, C. and KNIGHT, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71** 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

MARS, R. B., JBABDI, S. and RUSHWORTH, M. F. S. (2021). A common space approach to comparative neuroscience. *Annu. Rev. Neurosci.* **44** 69–86. https://doi.org/10.1146/annurev-neuro-100220-025942

MITRA, R. and ZHANG, C.-H. (2016). The benefit of group sparsity in group inference with de-biased scaled group Lasso. *Electron. J. Stat.* **10** 1829–1873. MR3522662 https://doi.org/10.1214/16-EJS1120

NEUHOUSER, M. L., SCHWARZ, Y., WANG, C., BREYMEYER, K., CORONADO, G., WANG, C.-Y., NOAR, K., SONG, X. and LAMPE, J. W. (2012). A low-glycemic load diet reduces serum C-reactive protein and modestly increases adiponectin in overweight and obese adults. *J. Nutr.* **142** 369–374.

NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. MR3611489 https://doi.org/10.1214/16-AOS1448

RANDOLPH, T. W., ZHAO, S., COPELAND, W., HULLAR, M. and SHOJAIE, A. (2018). Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.* **12** 540–566. MR3773404 https://doi.org/10.1214/17-AOAS1102

SCHAEFER, C. F., ANTHONY, K., KRUPA, S., BUCHOFF, J., DAY, M., HANNAY, T. and BUETOW, K. H. (2009). PID: The pathway interaction database. *Nucleic Acids Res.* **37** D674–D679. https://doi.org/10.1093/nar/gkn653

SEPICH-POORE, G. D., ZITVOGEL, L., STRAUSSMAN, R., HASTY, J., WARGO, J. A. and KNIGHT, R. (2021). The microbiome and human cancer. *Science* **371** eabc4552. https://doi.org/10.1126/science.abc4552

SHARIFI, F. and YE, Y. (2017). From gene annotation to function prediction for metagenomics. In *Protein Function Prediction* 27–34. Springer, Berlin.

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

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

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

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

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

WANG, Y., RANDOLPH, T. W., SHOJAIE, A. and MA, J. (2019). The generalized matrix decomposition biplot and its application to microbiome data. *mSystems* **4** e00504-19. https://doi.org/10.1128/mSystems.00504-19

WANG, Y., SHOJAIE, A., RANDOLPH, T., KNIGHT, P. and MA, J. (2023). Supplement to "Generalized matrix decomposition regression: Estimation and inference for two-way structured data." https://doi.org/10.1214/23-AOAS1746SUPPA, https://doi.org/10.1214/23-AOAS1746SUPPB

WASHBURNE, A. D., MORTON, J. T., SANDERS, J., MCDONALD, D., ZHU, Q., OLIVERIO, A. M. and KNIGHT, R. (2018). Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* **3** 652–661.

XU, Y., WANG, N., TAN, H.-Y., LI, S., ZHANG, C. and FENG, Y. (2020). Function of *Akkermansia muciniphila* in obesity: Interactions with lipid metabolism, immune response and gut systems. *Front. Microbiol.* **11** 219.

YATSUNENKO, T., REY, F. E., MANARY, M. J., TREHAN, I., DOMINGUEZ-BELLO, M. G., CONTRERAS, M., MAGRIS, M., HIDALGO, G., BALDASSANO, R. N. et al. (2012). Human gut microbiome viewed across age and geography. *Nature* **486** 222–227. https://doi.org/10.1038/nature11053

YU, G. and BIEN, J. (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika* **106** 533–546. MR3992388 https://doi.org/10.1093/biomet/asz017

ZEEVI, D., KOREM, T., GODNEVA, A., BAR, N., KURILSHIKOV, A., LOTAN-POMPAN, M., WEINBERGER, A., FU, J., WIJMENGA, C. et al. (2019). Structural variation in the gut microbiome associates with host health. *Nature* **568** 43–48.

ZHAN, X., PLANTINGA, A., ZHAO, N. and WU, M. C. (2017). A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics* **73** 1453–1463. MR3744557 https://doi.org/10.1111/biom.12684

ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

ZHANG, Y. and PAN, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: Competitors or complements? *Genet. Epidemiol.* **39** 149–155. https://doi.org/10.1002/gepi.21879

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

ZHANG, X., LI, L., BUTCHER, J., STINTZI, A. and FIGEYS, D. (2019). Advancing functional and translational microbiome research using meta-omics approaches. *Microbiome* **7** 1–12.

ZHAO, S. and SHOJAIE, A. (2016). A significance test for graph-constrained estimation. *Biometrics* **72** 484–493. MR3515775 https://doi.org/10.1111/biom.12418

ZHAO, N., CHEN, J., CARROLL, I. M., RINGEL-KULKA, T., EPSTEIN, M. P., ZHOU, H., ZHOU, J. J., RINGEL, Y., LI, H. et al. (2015). Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *Am. J. Hum. Genet.* **96** 797–807.

ZHU, Y. and BRADIC, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2970–2992
https://doi.org/10.1214/23-AOAS1747

TARGETING UNDERREPRESENTED POPULATIONS IN PRECISION MEDICINE: A FEDERATED TRANSFER LEARNING APPROACH

BY SAI LI[1,a], TIANXI CAI[2,b] AND RUI DUAN[2,c]

[1]*Institute of Statistics and Big Data, Renmin University of China,* [a]*saili@ruc.edu.cn*
[2]*Department of Biostatistics, Harvard T.H. Chan School of Public Health,* [b]*tcai@hsph.harvard.edu,* [c]*rduan@hsph.harvard.edu*

The limited representation of minorities and disadvantaged populations in large-scale clinical and genomics research poses a significant barrier to translating precision medicine research into practice. Prediction models are likely to underperform in underrepresented populations due to heterogeneity across populations, thereby exacerbating known health disparities. To address this issue, we propose FETA, a two-way data integration method that leverages a federated transfer learning approach to integrate heterogeneous data from diverse populations and multiple healthcare institutions, with a focus on a target population of interest having limited sample sizes. We show that FETA achieves performance comparable to the pooled analysis, where individual-level data is shared across institutions, with only a small number of communications across participating sites. Our theoretical analysis and simulation study demonstrate how FETA's estimation accuracy is influenced by communication budgets, privacy restrictions, and heterogeneity across populations. We apply FETA to multisite data from the electronic Medical Records and Genomics (eMERGE) Network to construct genetic risk prediction models for extreme obesity. Compared to models trained using target data only, source data only, and all data without accounting for population-level differences, FETA shows superior predictive performance. FETA has the potential to improve estimation and prediction accuracy in underrepresented populations and reduce the gap in model performance across populations.

REFERENCES

ASHLEY, E. A. (2016). Towards precision medicine. *Nat. Rev. Genet.* **17** 507–522.

BASTANI, H. (2020). Predicting with proxies: Transfer learning in high dimension. *Manage. Sci.* **67** 2657–3320.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

CAI, T., LIU, M. and XIA, Y. (2022). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *J. Amer. Statist. Assoc.* **117** 2105–2119. MR4528492 https://doi.org/10.1080/01621459.2021.1904958

CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949

CAI, M., XIAO, J., ZHANG, S. et al. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.* **108** 632–655.

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

COLLINS, R. (2012). What makes uk biobank special? *The Lancet* (*London, England*) **379** 1173–1174.

COLLINS, F. S. and VARMUS, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* **372** 793–795.

DUAN, R., NING, Y. and CHEN, Y. (2022). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika* **109** 67–83. MR4374641 https://doi.org/10.1093/biomet/asab007

DUAN, R., BOLAND, M. R., MOORE, J. H. and CHEN, Y. (2019). ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pacific Symposium on Biocomputing* 30–41.

Key words and phrases. Federated learning, health equity, precision medicine, risk prediction, transfer learning.

DUAN, R., LUO, C., SCHUEMIE, M. J. et al. (2020). Learning from local to global: An efficient distributed algorithm for modeling time-to-event data. *J. Amer. Med. Inform. Assoc.* **27** 1028–1036.

DUNCAN, L., SHEN, H., GELAYE, B., MEIJSEN, J., RESSLER, K., FELDMAN, M., PETERSON, R. and DOMINGUE, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10** 1–9.

GOLDEN, A. and KESSLER, C. (2020). Obesity and genetics. *J. Amer. Assoc. Nurse Pract.* **32** 493–496.

GOTTESMAN, O., KUIVANIEMI, H., TROMP, G., FAUCETT, W. A., LI, R., MANOLIO, T. A., SANDERSON, S. C., KANNRY, J., ZINBERG, R. et al. (2013). The electronic medical records and genomics (emerge) network: Past, present, and future. *Genet. Med.* **15** 761–771.

GUO, Z. (2020). Inference for high-dimensional maximin effects in heterogeneous regression models using a sampling approach. Preprint, arXiv:2011.07568.

JORDAN, M. I., LEE, J. D. and YANG, Y. (2019). Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.* **114** 668–681. MR3963171 https://doi.org/10.1080/01621459.2018.1429274

KAAMAN, M., RYDÉN, M., AXELSSON, T., NORDSTRÖM, E., SICARD, A., BOULOUMIE, A., LANGIN, D., ARNER, P. and DAHLMAN, I. (2006). Alox5ap expression, but not gene haplotypes, is associated with obesity and insulin resistance. *Int. J. Obes.* **30** 447–452.

KAPLAN, N. M. (1989). The deadly quartet: Upper-body obesity, glucose intolerance, hypertriglyceridemia, and hypertension. *Arch. Intern. Med.* **149** 1514–1520.

KRAFT, S. A., CHO, M. K., GILLESPIE, K. et al. (2018). Beyond consent: Building trusting relationships with diverse populations in precision medicine research. *Am. J. Bioethics* **18** 3–20.

KUSHIDA, C. A., NICHOLS, D. A., JADRNICEK, R. et al. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50** S82.

LAM, M., CHEN, C.-Y., LI, Z. et al. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat. Genet.* **51** 1670–1678.

LANDRY, L. G., ALI, N., WILLIAMS, D. R. et al. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff.* **37** 780–785.

LANGE, K., PAPP, J. C., SINSHEIMER, J. S. and SOBEL, E. M. (2014). Next generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. *Annu. Rev. Stat. Appl.* **1** 279.

LECUÉ, G. and RIGOLLET, P. (2014). Optimal learning with Q-aggregation. *Ann. Statist.* **42** 211–224. MR3178462 https://doi.org/10.1214/13-AOS1190

LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5, 30. MR3625709

LI, S., CAI, T. and DUAN, R. (2023). Supplement to "Targeting underrepresented populations in precision medicine: A federated transfer learning approach." https://doi.org/10.1214/23-AOAS1747SUPP

LI, S., CAI, T. T. and LI, H. (2020). Transfer learning in large-scale gaussian graphical models with false discovery rate control.

LI, S., CAI, T. T. and LI, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 149–173. MR4400393

LI, R., LIN, D. K. J. and LI, B. (2013). Statistical inference in massive data sets. *Appl. Stoch. Models Bus. Ind.* **29** 399–409. MR3117826 https://doi.org/10.1002/asmb.1927

LI, R., CHEN, Y., RITCHIE, M. D. et al. (2020). Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21** 493–502.

LIU, M., XIA, Y., CHO, K. and CAI, T. (2021). Integrative high dimensional multiple testing with heterogeneity under data sharing constraints. *J. Mach. Learn. Res.* **22** Paper No. 126, 26. MR4279777

LOOS, R. J. and YEO, G. S. (2022). The genetics of obesity: From discovery to biology. *Nat. Rev. Genet.* **23** 120–133.

MARTIN, A. R., KANAI, M., KAMATANI, Y. et al. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51** 584–591.

MCCARTY, C. A., CHISHOLM, R. L., CHUTE, C. G. et al. (2011). The eMERGE network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genom.* **4** 1–11.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K. and GALSTYAN, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54** 1–35.

MORIN, O., VALLIÈRES, M., BRAUNSTEIN, S., GINART, J. B., UPADHAYA, T., WOODRUFF, H. C., ZWANENBURG, A., CHATTERJEE, A., VILLANUEVA-MEYER, J. E. et al. (2021). An artificial intelligence framework integrating longitudinal electronic health records with real-world data enables continuous pan-cancer prognostication. *Nat. Cancer* **2** 709–722.

PAN, L., FREEDMAN, D. S., GILLESPIE, C., PARK, S. and SHERRY, B. (2011). Incidences of obesity and extreme obesity among us adults: Findings from the 2009 behavioral risk factor surveillance system. *Popul. Health Metr.* **9** 1–9.

QIAN, J., TANIGAWA, Y., DU, W. et al. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet*. **16** e1009141.

RATHOD, R., ZHANG, H., KARMAUS, W., EWART, S., MZAYEK, F., ARSHAD, S. H. and HOLLOWAY, J. W. (2022). Association of childhood bmi trajectory with post-adolescent and adult lung function is mediated by pre-adolescent dna methylation. *Respir. Res*. **23** 1–11.

RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist*. **39** 731–771. MR2816337 https://doi.org/10.1214/10-AOS854

SANKAR, P. L. and PARKER, L. S. (2017). The precision medicine initiative's all of us research program: An agenda for research on its ethical, legal, and social issues. *Genet. Med*. **19** 743–750.

STOLPE, M., BHADURI, K. and DAS, K. (2016). Distributed support vector machines: An overview. In *Solving Large Scale Learning Tasks. Lecture Notes in Computer Science* **9580** 109–138. Springer, Cham. MR3537495 https://doi.org/10.1007/978-3-319-41706-6_5

SUDLOW, C., GALLACHER, J., ALLEN, N. et al. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. **12** e1001779.

TIAN, Y. and FENG, Y. (2022). Transfer Learning under High-dimensional Generalized Linear Models. *J. Amer. Statist. Assoc*. **0** 1-14. https://doi.org/10.1080/01621459.2022.2071278

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

TSYBAKOV, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. IV* 225–246. Kyung Moon Sa, Seoul. MR3727610

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist*. **36** 614–645. MR2396809 https://doi.org/10.1214/009053607000000929

VAN DER HAAK, M., WOLFF, A. C., BRANDNER, R. et al. (2003). Data security and protection in cross-institutional electronic patient records. *Int. J. Med. Inform*. **70** 117–130.

WALFORD, G. A., GUSTAFSSON, S., RYBIN, D., STANČÁKOVÁ, A., CHEN, H., LIU, C.-T., HONG, J., JENSEN, R. A., RICE, K. et al. (2016). Genome-wide association study of the modified stumvoll insulin sensitivity index identifies bcl2 and fam19a2 as novel insulin sensitivity loci. *Diabetes* **65** 3200–3211.

WANG, Y., O'CONNELL, J. R., MCARDLE, P. F., WADE, J. B., DORFF, S. E., SHAH, S. J., SHI, X., PAN, L., RAMPERSAUD, E. et al. (2009). Whole-genome association study identifies stk39 as a hypertension susceptibility gene. *Proc. Natl. Acad. Sci. USA* **106** 226–231.

WANG, X., YANG, Z., CHEN, X. and LIU, W. (2019a). Distributed inference for linear support vector machine. *J. Mach. Learn. Res*. **20** Paper No. 113, 41. MR3990467

WANG, Y., SONG, H., WANG, W. and ZHANG, Z. (2019b). Generation and characterization of megf6 null and cre knock-in alleles. *Genesis* **57** e23262.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 1–40.

WEST, K. M., BLACKSHER, E. and BURKE, W. (2017). Genomics, health disparities, and missed opportunities for the nation's research agenda. *JAMA* **317** 1831–1832.

WU, J., ROY, J. and STEWART, W. F. (2010). Prediction modeling using ehr data: Challenges, strategies, and a comparison of machine learning approaches. Medical care. S106–S113.

ZHOU, W. et al. (2021). Global biobank meta-analysis initiative: Powering genetic discovery across human diseases. MedRxiv.

ZILLIKENS, M. C., DEMISSIE, S., HSU, Y.-H., YERGES-ARMSTRONG, L. M., CHOU, W.-C., STOLK, L., LIVSHITS, G., BROER, L., JOHNSON, T. et al. (2017). Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun*. **8** 1–13.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2993–3012
https://doi.org/10.1214/23-AOAS1748

BUILDING A DOSE TOXO-EQUIVALENCE MODEL FROM A BAYESIAN META-ANALYSIS OF PUBLISHED CLINICAL TRIALS

BY ELIZABETH A. SIGWORTH[1,a], SAMUEL M. RUBINSTEIN[2,c], JEREMY L. WARNER[3,d], YONG CHEN[4,e] AND QINGXIA CHEN[1,b]

[1]*Department of Biostatistics, Vanderbilt University,* [a]*elizabeth.a.sigworth@vanderbilt.edu,* [b]*cindy.chen@vumc.org*
[2]*Division of Hematology, University of North Carolina School of Medicine,* [c]*samuel_rubinstein@med.unc.edu*
[3]*Department of Medicine, Vanderbilt University School of Medicine,* [d]*jeremy.warner@vumc.org*
[4]*Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania,* [e]*ychen123@pennmedicine.upenn.edu*

In clinical practice medications are often interchanged in treatment protocols when a patient negatively reacts to their first line of therapy. Although switching between medications is common, clinicians often lack structured guidance when choosing the initial dose and frequency of a new medication, given the former with respect to risk of adverse events. In this paper we propose to establish this dose toxo-equivalence relationship using published clinical trial results with one or both drugs of interest via a Bayesian meta-analysis model that accounts for both within- and between-study variances. With the posterior parameter samples from this model, we compute median and 95% credible intervals for equivalent dose pairs of the two drugs that are predicted to produce equal rates of an adverse outcome, relying solely on study-level information. Via extensive simulations, we show that this approach approximates well the true dose toxo-equivalence relationship, considering different study designs, levels of between-study variance, and the inclusion/exclusion of nonconfounder/nonmodifier subject-level covariates in addition to study-level covariates. We compare the performance of this study-level meta-analysis estimate to the equivalent individual patient data meta-analysis model and find comparable bias and minimal efficiency loss in the study-level coefficients used in the dose toxo-equivalence relationship. Finally, we present the findings of our dose toxo-equivalence model applied to two chemotherapy drugs, based on data from 169 published clinical trials.

REFERENCES

ARGYRIOU, A. A., KOLTZENBURG, M., POLYCHRONOPOULOS, P., PAPAPETROPOULOS, S. and KALOFONOS, H. P. (2008). Peripheral nerve damage associated with administration of taxanes in patients with cancer. *Crit. Rev. Oncol. Hematol.* **66** 218–228. https://doi.org/10.1016/j.critrevonc.2008.01.008

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. MR1702200 https://doi.org/10.1214/ss/1009211803

BERLIN, J. A., SANTANNA, J., SCHMID, C. H., SZCZECH, L. A., FELDMAN, H. I. and ANTI-LYMPHOCYTE ANTIBODY INDUCTION THERAPY STUDY GROUP (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Stat. Med.* **21** 371–387. https://doi.org/10.1002/sim.1023

CAI, T., PARAST, L. and RYAN, L. (2010). Meta-analysis for rare events. *Stat. Med.* **29** 2078–2089. MR2756556 https://doi.org/10.1002/sim.3964

CATES, C. J. (2002). Simpson's paradox and calculation of number needed to treat from meta-analysis. *BMC Med. Res. Methodol.* **2** 1. https://doi.org/10.1186/1471-2288-2-1

DEBRAY, T. P., MOONS, K. G., VAN VALKENHOEF, G., EFTHIMIOU, O., HUMMEL, N., GROENWOLD, R. H., REITSMA, J. B. and GROUP, G. M. R. (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res. Synth. Methods* **6** 293–309.

Key words and phrases. Bayesian methods, dose toxo-equivalence, individual patient data meta-analysis, study-level meta-analysis.

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

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL. MR0798597

KRUSCHKE, J. K. (2015). *Doing Bayesian Data Analysis*: *A Tutorial with R*, *Jags*, *and Stan*. Academic Press, San Diego.

LAMBERT, P. C., SUTTON, A. J., ABRAMS, K. R. and JONES, D. R. (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J. Clin. Epidemiol*. **55** 86–94.

LEBOVITS, A. H., STRAIN, J. J., MESSE, M. R., SCHLEIFER, S. J., TANAKA, J. S. and BHARDWAJ, S. (1990). Patient noncompliance with self-administered chemotherapy. *Cancer* **65** 17–22.

LIN, D.-Y. and ZENG, D. (2010a). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic Epidemiology*: *The Official Publication of the International Genetic Epidemiology Society* **34** 60–66.

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

LUO, C., ISLAM, M., SHEILS, N. E., BURESH, J., REPS, J., SCHUEMIE, M. J., RYAN, P. B., EDMONDSON, M., DUAN, R. et al. (2022). DLMM as a lossless one-shot algorithm for collaborative multi-site distributed linear mixed models. *Nat. Commun*. **13** 1–10.

LUTZ, W., LOWRY, J., KOPTA, S. M., EINSTEIN, D. A. and HOWARD, K. I. (2001). Prediction of dose-response relations based on patient characteristics. *J. Clin. Psychol*. **57** 889–900. https://doi.org/10.1002/jclp.1057

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

OLKIN, I. and SAMPSON, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics* **54** 317–322. MR1626809 https://doi.org/10.2307/2534018

SEVERINI, T. A. (1998). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85** 507–522. MR1665861 https://doi.org/10.1093/biomet/85.3.507

SIGWORTH, E. A., RUBINSTEIN, S. M., CHAUGAI, S., RIVERA, D. R., WALKER, P. D., CHEN, Q. and WARNER, J. L. (2022). Development of a Bayesian toxo-equivalence model between docetaxel and paclitaxel. *iScience* **25** 104045.

SIGWORTH, E. A., RUBINSTEIN, S. M., WARNER, J. L., CHEN, Y. and CHEN, Q. (2023). Supplement to "Building a dose toxo-equivalence model from a Bayesian meta-analysis of published clinical trials." https://doi.org/10.1214/23-AOAS1748SUPPA, https://doi.org/10.1214/23-AOAS1748SUPPB

STEINBERG, K., SMITH, S., STROUP, D., OLKIN, I., LEE, N., WILLIAMSON, G. and THACKER, S. (1997). Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Amer. J. Epidemiol*. **145** 917–925.

WARNER, J. L., COWAN, A. J., HALL, A. C. and YANG, P. C. (2015). HemOnc.org: A collaborative online knowledge platform for oncology professionals. *Journal of Oncology Practice* **11** e336–e350.

WHITEHEAD, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Wiley, New York.

ZENG, D. and LIN, D. Y. (2015). On random-effects meta-analysis. *Biometrika* **102** 281–294. MR3371004 https://doi.org/10.1093/biomet/asv011

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3013–3034
https://doi.org/10.1214/23-AOAS1749

A BAYESIAN GROUP SELECTION WITH COMPOSITIONAL RESPONSES FOR ANALYSIS OF RADIOLOGIC TUMOR PROPORTIONS AND THEIR GENOMIC DETERMINANTS

BY THIERRY CHEKOUO[1,a], FRANCESCO C. STINGO[2,b], SHARIQ MOHAMMED[3,c], ARVIND RAO[4,d] AND VEERABHADRAN BALADANDAYUTHAPANI[5,e]

[1]*Division of Biostatistics, University of Minnesota,* [a]*tchekouo@umn.edu*
[2]*Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence,* [b]*francescoclaudio.stingo@unifi.it*
[3]*Department of Biostatistics, Boston University,* [c]*shariqm@bu.edu*
[4]*Department of Computational Medicine & Bioinformatics, University of Michigan,* [d]*ukarvind@umich.edu*
[5]*Department of Biostatistics, University of Michigan,* [e]*veerab@umich.edu*

Volumetric imaging features are used in cancer research to determine the size and the composition of a tumor and have been shown to be prognostic of overall survival. In this paper we focus on the analysis of tumor component proportions of brain cancer patients collected through The Cancer Genome Atlas (TCGA) project. Our main goal is to identify pathways and corresponding genes that can explain the heterogeneity of the composition of a brain tumor. In particular, we focus on the glioblastoma multiform (GBM), as it is the most common malignant brain neoplasm, accounting for 23% of all primary brain tumors for which it still has very poor prognosis. We propose a Bayesian hierarchical model for variable selection with a group structure in the context of correlated multivariate compositional response variables. More specifically, we model the proportions of the tumor components within the tumor using a Dirichlet model by allowing for straightforward incorporation of available high-dimensional covariate information within a log-linear regression framework. We impose prior distributions that account for the overlapping structure between groups of covariates. Simulations and application to GBM disease show the importance of our approach. We have identified associations between tumor component volume-based features and several important pathways and genes. Some of these genes have previously been shown to be prognostic indicators of overall survival time in GBM.

REFERENCES

AITCHISON, J. (1982). The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B* **44** 139–177. MR0676206

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability*. CRC Press, London. MR0865647 https://doi.org/10.1007/978-94-009-4109-0

BARCELÓ, C., PAWLOWSKY, V. and GRUNSKY, E. (1996). Some aspects of transformations of compositional data and the identification of outliers. *Math. Geol.* **28** 501–518.

BARRETT, T., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M. et al. (2012). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Res.* **41** D991–D995.

CAMARGO, A. P., STERN, J. M. and LAURETTO, M. S. (2012). Estimation and model selection in Dirichlet regression. *AIP Conf. Proc.* **1443** 206–213.

CARLIN, B. and CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57** 473–484.

CASELLA, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics* **2** 485–500. https://doi.org/10.1093/biostatistics/2.4.485

Key words and phrases. Glioblastoma, Dirichlet regression, Bayesian hierarchical model, group selection, overlap.

CECCARELLI, M., BARTHEL, F. P., MALTA, T. M., SABEDOT, T. S., SALAMA, S. R., MURRAY, B. A., MOROZOVA, O., NEWTON, Y., RADENBAUGH, A. et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164** 550–563.

CHEKOUO, T., MOHAMMED, S. and RAO, A. (2020). A Bayesian 2D functional linear model for gray-level co-occurrence matrices in texture analysis of lower grade gliomas. *NeuroImage Clin.* **28** 102437. https://doi.org/10.1016/j.nicl.2020.102437

CHEKOUO, T. and SAFO, S. E. (2023). Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. *Biostatistics* **24** 124–139. MR4522707 https://doi.org/10.1093/biostatistics/kxab016

CHEKOUO, T., STINGO, F. C., DOECKE, J. D. and DO, K.-A. (2015). miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics* **71** 428–438. MR3366247 https://doi.org/10.1111/biom.12266

CHEKOUO, T., STINGO, F. C., GUINDANI, M. and DO, K.-A. (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. *Ann. Appl. Stat.* **10** 1547–1571. MR3553235 https://doi.org/10.1214/16-AOAS948

CHEKOUO, T., STINGO, F. C., DOECKE, J. D. and DO, K.-A. (2017). A Bayesian integrative approach for multi-platform genomic data: A kidney cancer case study. *Biometrics* **73** 615–624. MR3665977 https://doi.org/10.1111/biom.12587

CHEKOUO, T., STINGO, F. C., MOHAMMED, S., RAO, A. and BALADANDAYUTHAPANI, V. (2023). Supplement to "A Bayesian group selection with compositional responses for analysis of radiologic tumor proportions and their genomic determinants." https://doi.org/10.1214/23-AOAS1749SUPP

CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

CHEN, R.-B., CHU, C.-H., YUAN, S. and WU, Y. N. (2016). Bayesian sparse group selection. *J. Comput. Graph. Statist.* **25** 665–683. MR3533632 https://doi.org/10.1080/10618600.2015.1041636

COLEN, R. R., WANG, J., SINGH, S. K., GUTMAN, D. A. and ZINN, P. O. (2015). Glioblastoma: Imaging genomic mapping reveals sex-specific oncogenic associations of cell death. *Radiology* **275** 215–227.

CRESPO, S., KIND, M. and ARCARO, A. (2016). The role of the PI3K/AKT/mTOR pathway in brain tumor metastasis. *J. Cancer Metastasis Treat.* **2** 80–89.

DIEHN, M., NARDINI, C., WANG, D. S., MCGOVERN, S., JAYARAMAN, M., LIANG, Y., ALDAPE, K., CHA, S. and KUO, M. D. (2008). Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. USA* **105** 5213–5218.

DING, G., ZHOU, L., SHEN, T. and CAO, L. (2018). IFN-γ induces the upregulation of RFXAP via inhibition of miR-212-3p in pancreatic cancer cells: A novel mechanism for IFN-γ response. *Oncol. Lett.* **15** 3760–3765.

GROSSMANN, P., GUTMAN, D. A., DUNN, W. D., HOLDER, C. A. and AERTS, H. J. W. L. (2016). Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in glioblastoma. *BMC Cancer* **16** 611.

GUTMAN, D. A., DUNN, W. D., GROSSMANN, P., COOPER, L. A. D., HOLDER, C. A., LIGON, K. L., ALEXANDER, B. M. and AERTS, H. J. W. L. (2015). Somatic mutations associated with MRI-derived volumetric features in glioblastoma. *Neuroradiology* **57** 1227–1237.

HARALICK, R. M., SHANMUGAM, K. and DINSTEIN, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3** 610–621.

HIJAZI, R. H. and JERNIGAN, R. W. (2009). Modeling compositional data using Dirichlet regression models. *J. Appl. Probab. Stat.* **4** 77–91. MR2668780

HUANG, J., BREHENY, P. and MA, S. (2012). A selective review of group selection in high-dimensional models. *Statist. Sci.* **27** 481–499. MR3025130 https://doi.org/10.1214/12-STS392

HUANG, C., CHEN, D., ZHU, H., LV, S., LI, Q. and LI, G. (2019). LITAF enhances radiosensitivity of human glioma cells via the FoxO1 pathway. *Cell. Mol. Neurobiol.* **39** 871–882. https://doi.org/10.1007/s10571-019-00686-4

ILIADIS, G., KOTOULA, V., CHATZISOTIRIOU, A., TELEVANTOU, D., ELEFTHERAKI, A. G., LAMBAKI, S., MISAILIDOU, D., SELVIARIDIS, P. and FOUNTZILAS, G. (2012). Volumetric and MGMT parameters in glioblastoma patients: Survival analysis. *BMC Cancer* **12** 3–3.

IRSA (2017). International Radiosurgery Association. Available at http://www.irsa.org/glioblastoma.html. Accessed: 2017-11-20.

JAIN, R., POISSON, L. M., GUTMAN, D., SCARPACE, L., HWANG, S. N., HOLDER, C. A., WINTERMARK, M., RAO, A., COLEN, R. R. et al. (2014). Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: Focus on the nonenhancing component of the tumor. *Radiology* **272** 484–493.

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

KIM, H., HUANG, W., JIANG, X., PENNICOOKE, B., PARK, P. J. and JOHNSON, M. D. (2010). Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proc. Natl. Acad. Sci. USA* **107** 2183–2188.

KOTROTSOU, A., ZINN, P. O. and COLEN, R. R. (2016). Radiomics in brain tumors: An emerging technique for characterization of tumor environment. *Magn. Reson. Imaging Clin. N. Am.* **24** 719–729. https://doi.org/10.1016/j.mric.2016.06.006

LACROIX, M., ABI-SAID, D., FOURNEY, D. R., GOKASLAN, Z. L., SHI, W., DEMONTE, F., LANG, F. F., MCCUTCHEON, I. E., HASSENBUSCH, S. J. et al. (2001). A multivariate analysis of 416 patients with glioblastoma multiforme: Prognosis, extent of resection, and survival. *J. Neurosurg.* **95** 190–198.

LAI, W.-T. and CHEN, R.-B. (2021). A review of Bayesian group selection approaches for linear regression models. *Wiley Interdiscip. Rev.: Comput. Stat.* **13** Paper No. e1513, 22. MR4272068 https://doi.org/10.1002/wics.1513

LI, W. and CHEKOUO, T. (2022). Bayesian group selection with non-local priors. *Comput. Statist.* **37** 287–302. MR4390014 https://doi.org/10.1007/s00180-021-01115-1

LI, Y., NAN, B. and ZHU, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240 https://doi.org/10.1111/biom.12292

LI, Y., NAN, B. and ZHU, J. (2016). MSGLasso: Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. R package version 2.1.

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDÒTTIR, H., TAMAYO, P. and MESIROV, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinform.* **27** 1739–1740.

LIBERZON, A., BIRGER, C., THORVALDSDÓTTIR, H., GHANDI, M., MESIROV, J. P. and TAMAYO, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1** 417–425. https://doi.org/10.1016/j.cels.2015.12.004

LIQUET, B. and SUTTON, M. (2016). MBSGS: Multivariate Bayesian sparse group selection with spike and slab. R package version 1.0.0.

LIQUET, B., DE MICHEAUX, P. L., HEJBLUM, B. P. and THIÉBAUT, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* **32** 35–42. https://doi.org/10.1093/bioinformatics/btv535

LIQUET, B., MENGERSEN, K., PETTITT, A. N. and SUTTON, M. (2017). Bayesian variable selection regression of multivariate responses for group data. *Bayesian Anal.* **12** 1039–1067. MR3724978 https://doi.org/10.1214/17-BA1081

LIU, J., JI, S. and YE, J. (2009). Multi-task feature learning via efficient L2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. *UAI'*09 339–348. AUAI Press, Arlington, VA.

LIU, H., PALATUCCI, M. and ZHANG, J. (2009). Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *International Conference on Machine Learning*, Pittsburgh, PA.

MAIER, M. J. (2014). DirichletReg: Dirichlet regression for compositional data in R. Research Report Series/Dept. Statistics and Mathematics No. 125, WU Vienna Univ. Economics and Business, Vienna. Available at http://epub.wu.ac.at/4077/.

MAIER, M. J. (2015). DirichletReg: Dirichlet regression in R. R package version 0.6-3.

MAO, H., LEBRUN, D. G., YANG, J., ZHU, V. F. and LI, M. (2012). Deregulated signaling pathways in glioblastoma multiforme: Molecular mechanisms and therapeutic targets. *Cancer Investig.* **30** 48–56.

MEIER, L. (2015). grplasso: Fitting user specified models with group lasso penalty. R package version 0.4-5.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 53–71. MR2412631 https://doi.org/10.1111/j.1467-9868.2007.00627.x

MEIER, R., KNECHT, U., LOOSLI, T., BAUER, S., SLOTBOOM, J., WIEST, R. and REYES, M. (2016). Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Sci. Rep.* **6** 23376.

MOHAMMED, S., BHARATH, K., KURTEK, S., RAO, A. and BALADANDAYUTHAPANI, V. (2021). RADIOHEAD: Radiogenomic analysis incorporating tumor heterogeneity in imaging through densities. *Ann. Appl. Stat.* **15** 1808–1830. MR4355077 https://doi.org/10.1214/21-aoas1458

MOSIMANN, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299 https://doi.org/10.1093/biomet/49.1-2.65

NAEINI, K. M., POPE, W. B., CLOUGHESY, T. F., HARRIS, R. J., LAI, A., ESKIN, A., CHOWDHURY, R., PHILLIPS, H. S., NGHIEMPHU, P. L. et al. (2013). Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. *Neuro-Oncol.* **15** 626–634.

NARANG, S., LEHRER, M., YANG, D., LEE, J. and RAO, A. (2016). Radiomics in glioblastoma: Current status, challenges and potential opportunities. *Transl. Cancer Res.* **5**.

NARANG, S., KIM, D., AITHALA, S., HEIMBERGER, A. B., AHMED, S., RAO, D., RAO, G. and RAO, A. (2017). Tumor image-derived texture features are associated with CD3 T-cell infiltration status in glioblastoma. *Oncotarget* **8** 101244–101254. https://doi.org/10.18632/oncotarget.20643

ORMEROD, J. T., YOU, C. and MÜLLER, S. (2017). A variational Bayes approach to variable selection. *Electron. J. Stat.* **11** 3549–3594. MR3709863 https://doi.org/10.1214/17-EJS1332

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103** 681–686. MR2524001 https://doi.org/10.1198/016214508000000337

PAWLOWSKY-GLAHN, V., EGOZCUE, J. J. and TOLOSANA-DELGADO, R. (2015). *Modeling and Analysis of Compositional Data. Statistics in Practice*. Wiley, Chichester. MR3328965

POPE, W. B., SAYRE, J., PERLINA, A., VILLABLANCA, J. P., MISCHEL, P. S. and CLOUGHESY, T. F. (2005). MR imaging correlates of survival in patients with high-grade gliomas. *AJNR Am. J. Neuroradiol.* **26** 2466–2474.

RAHMIM, A., SCHMIDTLEIN, C. R., JACKSON, A. J., SHEIKHBAHAEI, S., MARCUS, C. V., ASHRAFINIA, S., SOLTANI, M. and SUBRAMANIAM, R. M. (2016). A novel metric for quantification of homogeneous and heterogeneous tumors in PET for enhanced clinical outcome prediction. *Phys. Med. Biol.* **61** 227–42.

RAMAN, S., FUCHS, T. J., WILD, P. J., DAHL, E. and ROTH, V. (2009). The Bayesian group-lasso for analyzing contingency tables. In *Proceedings of the* 26*th Annual International Conference on Machine Learning. ICML'*09 881–888. ACM, New York, NY. https://doi.org/10.1145/1553374.1553487

RAY, K. and SZABÓ, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *J. Am. Statist. Assoc.* **117** 1270–1281. MR4480711 https://doi.org/10.1080/01621459.2020.1847121

RAYENS, W. S. and SRINIVASAN, C. (1991). Estimation in compositional data analysis. *J. Chemom.* **5** 361–374.

ROCKOVA, V. and LESAFFRE, E. (2014). Incorporating grouping information in Bayesian variable selection with applications in genomics. *Bayesian Anal.* **9** 221–258. MR3188306 https://doi.org/10.1214/13-BA846

RONALD, W. and YAIR, G. (2018 (accessed August, 2020)). Diffuse astrocytoma. Available at http://mayfieldclinic.com/pe-braintumor.htm.

SANCHEZ-VEGA, F., MINA, M., ARMENIA, J., CHATILA, W. K., LUNA, A., LA, K. C., DIMITRIADOY, S., LIU, D. L., KANTHETI, H. S. et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173** 321–337.e10.

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

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

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013b). SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization. R package version 1.1.

STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929 https://doi.org/10.1214/11-AOAS463

SUBRAHMANYA, N. and SHIN, Y. C. (2013). A variational Bayesian framework for group feature selection. *Int. J. Mach. Learn. Cybern.* **4** 609–619.

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

THIBAULT, G., FERTIL, B., NAVARRO, C., PEREIRA, S., CAU, P., LEVY, N., SEQUEIRA, J. and MARI, J.-L. (2013). Shape and texture indexes application to cell nuclei classification. *Int. J. Pattern Recognit. Artif. Intell.* **27** 1357002, 23. MR3046234 https://doi.org/10.1142/S0218001413570024

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

VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *J. Am. Statist. Assoc.* **103** 790–796. MR2524010 https://doi.org/10.1198/016214508000000409

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

WADSWORTH, W. D., ARGIENTO, R., GUINDANI, M., GALLOWAY-PENA, J., SHELBURNE, S. A. and VANNUCCI, M. (2017). An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinform.* **18** 94. https://doi.org/10.1186/s12859-017-1516-0

WANG, S., NAN, B., ZHOU, N. and ZHU, J. (2009). Hierarchically penalized Cox regression with grouped variables. *Biometrika* **96** 307–322. MR2507145 https://doi.org/10.1093/biomet/asp016

WEN, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70** 73–83. MR3251668 https://doi.org/10.1111/biom.12112

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

YAMASHITA, D., KONDO, T., OHUE, S., TAKAHASHI, H., ISHIKAWA, M., MATOBA, R., SUEHIRO, S., KOHNO, S., HARADA, H. et al. (2015). miR340 suppresses the stem-like cell function of glioma-initiating cells by targeting tissue plasminogen activator. *Cancer Res.* **75** 1123–1133.

YANG, X. and NARISETTY, N. N. (2020). Consistent group selection with Bayesian high dimensional modeling. *Bayesian Anal.* **15** 909–935. MR4132654 https://doi.org/10.1214/19-BA1178

YAVORSKI, J. M. and BLANCK, G. (2017). MHC class II associated stomach cancer mutations correlate with lack of subsequent tumor development. *Mol. Clin. Oncol.* **7** 1119–1121. https://doi.org/10.3892/mco.2017.1432

YIP, S. S. F. and AERTS, H. J. W. L. (2016). Applications and limitations of radiomics. *Phys. Med. Biol.* **61** R150.

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

ZENG, Y. and BREHENY, P. (2016). Overlapping group logistic regression with applications to genetic pathway selection. *Cancer Inform.* **15** 179–187. https://doi.org/10.4137/CIN.S40043

ZHAI, L., SPRANGER, S., BINDER, D. C., GRITSINA, G., LAUING, K. L., GILES, F. J. and WAINWRIGHT, D. A. (2015). Molecular pathways: Targeting IDO1 and other tryptophan dioxygenases for cancer immunotherapy. *Clin. Cancer Res.* **21** 5427–5433. https://doi.org/10.1158/1078-0432.CCR-15-0420

ZHANG, J. D. and WIEMANN, S. (2009). KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25** 1470–1471.

ZHANG, L., BALADANDAYUTHAPANI, V., MALLICK, B. K., MANYAM, G. C., THOMPSON, P. A., BONDY, M. L. and DO, K.-A. (2014a). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 595–620. MR3258055 https://doi.org/10.1111/rssc.12053

ZHANG, L., MORRIS, J. S., ZHANG, J., ORLOWSKI, R. Z. and BALADANDAYUTHAPANI, V. (2014b). Bayesian joint selection of genes and pathways: Applications in multiple myeloma genomics. *Cancer Inform.* **13** 113–123.

ZHU, L., HUO, Z., MA, T., OESTERREICH, S. and TSENG, G. C. (2019). Bayesian indicator variable selection to incorporate hierarchical overlapping group structure in multi-omics applications. *Ann. Appl. Stat.* **13** 2611–2636. MR4037443 https://doi.org/10.1214/19-aoas1271

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3035–3055
https://doi.org/10.1214/23-AOAS1750

A BAYESIAN DECISION FRAMEWORK FOR OPTIMIZING SEQUENTIAL COMBINATION ANTIRETROVIRAL THERAPY IN PEOPLE WITH HIV

BY WEI JIN[1,a], YANG NI[2,c], JANE O'HALLORAN[3,d], AMANDA B. SPENCE[4,e], LEAH H. RUBIN[5,f] AND YANXUN XU[1,b]

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

Numerous adverse effects (e.g., depression) have been reported for combination antiretroviral therapy (cART) despite its remarkable success in viral suppression in people with HIV (PWH). To improve long-term health outcomes for PWH, there is an urgent need to design personalized optimal cART with the lowest risk of comorbidity in the emerging field of precision medicine for HIV. Large-scale HIV studies offer researchers unprecedented opportunities to optimize personalized cART in a data-driven manner. However, the large number of possible drug combinations for cART makes the estimation of cART effects a high-dimensional combinatorial problem, imposing challenges in both statistical inference and decision-making. We develop a two-step Bayesian decision framework for optimizing sequential cART assignments. In the first step, we propose a dynamic model for individuals' longitudinal observations using a multivariate Gaussian process. In the second step, we build a probabilistic generative model for cART assignments and design an uncertainty-penalized policy optimization using the uncertainty quantification from the first step. Applying the proposed method to a dataset from the Women's Interagency HIV Study, we demonstrate its clinical utility in assisting physicians to make effective treatment decisions, serving the purpose of both viral suppression and comorbidity risk reduction.

REFERENCES

ADIMORA, A. A., RAMIREZ, C., BENNING, L., GREENBLATT, R. M., KEMPF, M.-C. et al. (2018). Cohort profile: The women's interagency HIV study (WIHS). *Int. J. Epidemiol.* **47** 393–394i.

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

ALVAREZ, M. A., ROSASCO, L., LAWRENCE, N. D. et al. (2012). Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.* **4** 195–266.

ARENAS-PINTO, A., THOMPSON, J., MUSORO, G., MUSANA, H., LUGEMWA, A. et al. (2016). Peripheral neuropathy in HIV patients in sub-Saharan Africa failing first-line therapy and the response to second-line ART in the EARNEST trial. *J. Neurovirology* **22** 104–113.

BECK, D., COHN, T., HARDMEIER, C. and SPECIA, L. (2015). Learning structural kernels for natural language processing. *Trans. Assoc. Comput. Linguist.* **3** 461–473.

BLOOMFIELD, G. S., HOGAN, J. W., KETER, A., SANG, E., CARTER, E. J. et al. (2011). Hypertension and obesity as cardiovascular risk factors among HIV seropositive patients in Western Kenya. *PLoS ONE* **6** e22288.

BOGOJESKA, J., BICKEL, S., ALTMANN, A. and LENGAUER, T. (2010). Dealing with sparse data in predicting outcomes of HIV combination therapies. *Bioinformatics* **26** 2085–2092. https://doi.org/10.1093/bioinformatics/btq361

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

Key words and phrases. Antiretroviral therapy, multivariate Gaussian process, offline reinforcement learning, precision medicine, uncertainty-penalized policy optimization.

CHECA, A., CASTILLO, A., CAMACHO, M., TAPIA, W., HERNANDEZ, I. and TERAN, E. (2020). Depression is associated with efavirenz-containing treatments in newly antiretroviral therapy initiated HIV patients in Ecuador. *AIDS Res. Ther.* **17** 1–5.

D'SOUZA, G., GOLUB, E. T. and GANGE, S. J. (2019). The changing science of HIV epidemiology in the United States. *Amer. J. Epidemiol.* **188** 2061–2068. https://doi.org/10.1093/aje/kwz211

DIETRICH, L. G., THORBALL, C. W., RYOM, L., BURKHALTER, F., HASSE, B. et al. (2021). Rapid progression of kidney dysfunction in people living with HIV: Use of polygenic and data collection on adverse events of anti-HIV drugs (D: A: D) risk scores. *J. Infect. Dis.* **223** 2145–2153.

FUJIMOTO, S., MEGER, D. and PRECUP, D. (2019). Off-policy deep reinforcement learning without exploration. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 2052–2062. PMLR, Long Beach, CA.

GREENSMITH, E., BARTLETT, P. L. and BAXTER, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.* **5** 1471–1530. MR2248025 https://doi.org/10.1162/jmlr.2003.4.7-8.1471

HUA, W., MEI, H., ZOHAR, S., GIRAL, M. and XU, Y. (2022). Personalized dynamic treatment regimes in continuous time: A Bayesian approach for optimizing clinical decisions with timing. *Bayesian Anal.* **17** 849–878. MR4483241 https://doi.org/10.1214/21-ba1276

ISLAM, F. M., WU, J., JANSSON, J. and WILSON, D. P. (2012). Relative risk of renal disease among people living with HIV: A systematic review and meta-analysis. *BMC Public Health* **12** 1–15.

JIN, W., NI, Y., RUBIN, L. H., SPENCE, A. B. and XU, Y. (2022). A Bayesian nonparametric approach for inferring drug combination effects on mental health in people with HIV. *Biometrics* **78** 988–1000. MR4493503 https://doi.org/10.1111/biom.13508

JIN, W., NI, Y., O'HALLORAN, J., SPENCE, A. B., RUBIN, L. H. and XU, Y. (2023). Supplement to "A Bayesian decision framework for optimizing sequential combination antiretroviral therapy in people with HIV." https://doi.org/10.1214/23-AOAS1750SUPP

KENDALL, M. G. (1957). *A Course in Multivariate Analysis. Griffin's Statistical Monographs & Courses* **2**. Hafner, New York. MR0092297

LANGE, S., GABEL, T. and RIEDMILLER, M. (2012). Batch reinforcement learning. In *Reinforcement Learning* 45–73. Springer, Berlin.

LANGEBEEK, N., KOOIJ, K. W., WIT, F. W., STOLTE, I. G., SPRANGERS, M. A. et al. (2017). Impact of comorbidity and ageing on health-related quality of life in HIV-positive and HIV-negative individuals. *AIDS* **31** 1471–1481.

LEDERGERBER, B., EGGER, M., OPRAVIL, M., TELENTI, A., HIRSCHEL, B. et al. (1999). Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: A prospective cohort study. *Lancet* **353** 863–868.

LUNDGREN, J. D., MOCROFT, A., GATELL, J. M., LEDERGERBER, B., MONFORTE, A. D., HERMANS, P. et al. (2002). A clinically prognostic scoring system for patients receiving highly active antiretroviral therapy: Results from the EuroSIDA study. *J. Infect. Dis.* **185** 178–187.

MA, Q., VAIDA, F., WONG, J., SANDERS, C. A., KAO, Y. et al. (2016). Long-term efavirenz use is associated with worse neurocognitive functioning in HIV-infected patients. *J. Neurovirology* **22** 170–178.

MA, J., YANG, Q., HWANG, S.-J., FOX, C. S. and CHU, A. Y. (2017). Genetic risk score and risk of stage 3 chronic kidney disease. *BMC Nephrol.* **18** 1–6.

MASNOON, N., SHAKIB, S., KALISCH-ELLETT, L. and CAUGHEY, G. E. (2017). What is polypharmacy? A systematic review of definitions. *BMC Geriatr.* **17** 1–10.

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

POLLACK, T. M., DUONG, H. T., PHAM, T. T., DO, C. D. and COLBY, D. (2017). Cigarette smoking is associated with high HIV viral load among adults presenting for antiretroviral therapy in Vietnam. *PLoS ONE* **12** e0173534. https://doi.org/10.1371/journal.pone.0173534

RABOUD, J. M., MONTANER, J. S., CONWAY, B., RAE, S., REISS, P. et al. (1998). Suppression of plasma viral load below 20 copies/ml is required to achieve a long-term response to therapy. *AIDS* **12** 1619–1624.

RADLOFF, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Appl. Psychol. Meas.* **1** 385–401.

RIEDMILLER, M. (2005). Neural fitted Q iteration—first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning* 317–328. Springer, Berlin.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

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

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect*. *Notes Stat*. **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11

SUN, Y. and WANG, L. (2021). Stochastic tree search for estimating optimal dynamic treatment regimes. *J*. *Amer*. *Statist*. *Assoc*. **116** 421–432. MR4227704 https://doi.org/10.1080/01621459.2020.1819294

SURIAL, B., LEDERGERBER, B., CALMY, A., CAVASSINI, M., GÜNTHARD, H. F. et al. (2020). Changes in renal function after switching from TDF to TAF in HIV-infected individuals: A prospective cohort study. *J*. *Infect*. *Dis*. **222** 637–645.

WALLENIUS, K. T. (1963). Biased sampling; the noncentral hypergeometric probability distribution. Technical report, Stanford Univ. Applied Mathematics and Statistics Labs. MR2614179

WANG, L., ROTNITZKY, A., LIN, X., MILLIKAN, R. E. and THALL, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J*. *Amer*. *Statist*. *Assoc*. **107** 493–508. MR2980060 https://doi.org/10.1080/01621459.2011.641416

WEISSER, B., PREDEL, H.-G., GILLESSEN, A., HACKE, C., VOR DEM ESCHE, J. et al. (2020). Single pill regimen leads to better adherence and clinical outcome in daily practice in patients suffering from hypertension and/or dyslipidemia: Results of a meta-analysis. *High Blood Press*. *Cardiovasc*. *Prev*. **27** 157.

WILLIAMS, D. W., LI, Y., DASTGHEYB, R., FITZGERALD, K. C., MAKI, P. M. et al. (2020). Associations between antiretroviral drugs on depressive symptomatology in homogenous subgroups of women with HIV. *J*. *Neuroimmune Pharmacol*. 1–14.

XU, Y., MÜLLER, P., WAHED, A. S. and THALL, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J*. *Amer*. *Statist*. *Assoc*. **111** 921–950. MR3561917 https://doi.org/10.1080/01621459.2015.1086353

YAZDANPANAH, Y., SISSOKO, D., EGGER, M., MOUTON, Y., ZWAHLEN, M. et al. (2004). Clinical efficacy of antiretroviral combination therapy based on protease inhibitors or non-nucleoside analogue reverse transcriptase inhibitors: Indirect comparison of controlled trials. *BMJ* **328** 249.

YU, T., QUILLEN, D., HE, Z., JULIAN, R. et al. (2020a). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning* 1094–1100. PMLR, Osaka, Japan.

YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J. Y., LEVINE, S., FINN, C. and MA, T. (2020b). MOPO: Model-based offline policy optimization. *Adv*. *Neural Inf*. *Process*. *Syst*. **33** 14129–14142.

ZENG, C., GUO, Y., HONG, Y. A., GENTZ, S., ZHANG, J. et al. (2019). Differential effects of unemployment on depression in people living with HIV/AIDS: A quantile regression approach. *AIDS Care*.

ZICH, J. M., ATTKISSON, C. C. and GREENFIELD, T. K. (1990). Screening for depression in primary care clinics: The CES-D and the BDI. *Int*. *J*. *Psychiatr*. *Med*. **20** 259–277.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3056–3087
https://doi.org/10.1214/23-AOAS1751

A DYNAMIC SPATIAL FILTERING APPROACH TO MITIGATE UNDERESTIMATION BIAS IN FIELD CALIBRATED LOW-COST SENSOR AIR POLLUTION DATA

BY CLAIRE HEFFERNAN[1,a], ROGER PENG[2,c], DREW R. GENTNER[3,d], KIRSTEN KOEHLER[4,e] AND ABHIRUP DATTA[1,b]

[1]*Department of Biostatistics, Johns Hopkins University,* [a]*cmheff@jhu.edu,* [b]*abhidatta@jhu.edu*
[2]*Department of Statistics and Data Sciences, University of Texas, Austin,* [c]*roger.peng@austin.utexas.edu*
[3]*Department of Chemical & Environmental Engineering, Yale University,* [d]*drew.gentner@yale.edu*
[4]*Department of Environmental Health and Engineering, Johns Hopkins University,* [e]*kkoehle1@jhu.edu*

Low-cost air pollution sensors, offering hyperlocal characterization of pollutant concentrations, are becoming increasingly prevalent in environmental and public health research. However, low-cost air pollution data can be noisy, biased by environmental conditions, and usually need to be field-calibrated by collocating low-cost sensors with reference-grade instruments. We show, theoretically and empirically, that the common procedure of regression-based calibration, using collocated data, systematically underestimates high air pollution concentrations, which are critical to diagnose from a health perspective. Current calibration practices also often fail to utilize the spatial correlation in pollutant concentrations. We propose a novel spatial filtering approach to collocation-based calibration of low-cost networks that mitigates the underestimation issue by using an inverse regression. The inverse regression also allows for incorporating spatial correlations by a second-stage model for the true pollutant concentrations using a conditional Gaussian process. Our approach works with one or more collocated sites in the network and is dynamic, leveraging spatial correlation with the latest available reference data. Through extensive simulations, we demonstrate how the spatial filtering substantially improves estimation of pollutant concentrations and measures peak concentrations with greater accuracy. We apply the methodology for calibration of a low-cost $PM_{2.5}$ network in Baltimore, Maryland, and diagnose air pollution peaks that are missed by the regression-calibration.

REFERENCES

APTE, J. S., MESSIER, K. P., GANI, S., BRAUER, M., KIRCHSTETTER, T. W., LUNDEN, M. M., MARSHALL, J. D., PORTIER, C. J., VERMEULEN, R. C. H. et al. (2017). High-resolution air pollution mapping with Google street view cars: Exploiting big data. *Environ. Sci. Technol.* **51** 6999–7008. https://doi.org/10.1021/acs.est.7b00891

ARDON-DRYER, K., DRYER, Y., WILLIAMS, J. N. and MOGHIMI, N. (2020). Measurements of PM 2.5 with PurpleAir under atmospheric conditions. *Atmos. Meas. Tech.* **13** 5441–5458.

BALZANO, L. and NOWAK, R. (2007). Blind calibration of sensor networks. In *Proceedings of the* 6*th International Conference on Information Processing in Sensor Networks. IPSN'*07 79–88. Association for Computing Machinery, New York, NY, USA.

BARKJOHN, K. K., GANTT, B. and CLEMENTS, A. L. (2021). Development and application of a United States-wide correction for PM 2.5 data collected with the PurpleAir sensor. *Atmos. Meas. Tech.* **14** 4617–4637.

BI, J., WILDANI, A., CHANG, H. H. and LIU, Y. (2020). Incorporating low-cost sensor measurements into high-resolution PM2.5 modeling at a large spatial scale. *Environ. Sci. Technol.* **54** 2152–2162.

BIGI, A., MUELLER, M., GRANGE, S. K., GHERMANDI, G. and HUEGLIN, C. (2018). Performance of NO, NO2 low cost sensors and three calibration approaches within a real world application. *Atmos. Meas. Tech.* **11** 3717–3735.

Key words and phrases. Spatial statistics, Gaussian process, Bayesian, air pollution, low-cost sensors.

BUEHLER, C., XIONG, F., ZAMORA, M. L., SKOG, K. M., KOHRMAN-GLASER, J., COLTON, S., MCNAMARA, M., RYAN, K., REDLICH, C. et al. (2021). Stationary and portable multipollutant monitors for high-spatiotemporal-resolution air quality studies including online calibration. *Atmos. Meas. Tech.* **14** 995–1013.

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**.

CHADWICK, E., LE, K., PEI, Z., SAYAHI, T., RAPP, C., BUTTERFIELD, A. E. and KELLY, K. E. (2021). Technical note: Understanding the effect of Covid-19 on particle pollution using a low-cost sensor network. *J. Aerosol. Sci.* **155** 105766. https://doi.org/10.1016/j.jaerosci.2021.105766

CHOI, T. (2005). Posterior consistency in nonparametric regression problems under Gaussian process priors. PhD Thesis, Department of Statistics, Carnegie Mellon University.

CLOUGHERTY, J. E., KHEIRBEK, I., EISL, H. M., ROSS, Z., PEZESHKI, G., GORCZYNSKI, J. E., JOHNSON, S., MARKOWITZ, S., KASS, D. et al. (2013). Intra-urban spatial variability in wintertime street-level concentrations of multiple combustion-related air pollutants: The New York city community air survey (NYCCAS). *J. Expo. Sci. Environ. Epidemiol.* **23** 232–240.

CONSIDINE, E. M., REID, C. E., OGLETREE, M. R. and DYE, T. (2021). Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network. *Environ. Pollut.* **268** 115833. https://doi.org/10.1016/j.envpol.2020.115833

COWLES, M. K. and ZIMMERMAN, D. L. (2003). A Bayesian space-time analysis of acid deposition data combined from two monitoring networks. *J. Geophys. Res., Atmos.* **108**.

COWLES, M. K., ZIMMERMAN, D. L., CHRIST, A. and MCGINNIS, D. L. (2002). Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems. *J. Agric. Biol. Environ. Stat.* **7** 536–557.

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

DATTA, A., BANERJEE, S., FINLEY, A. O., HAMM, N. A. S. and SCHAAP, M. (2016b). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *Ann. Appl. Stat.* **10** 1286–1316. MR3553225 https://doi.org/10.1214/16-AOAS931

DATTA, A., SAHA, A., ZAMORA, M. L., BUEHLER, C., HAO, L., XIONG, F., GENTNER, D. R. and KOEHLER, K. (2020). Statistical field calibration of a low-cost PM. *Atmos. Environ.* **242**. https://doi.org/10.1016/j.atmosenv.2020.117761

DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–425. MR1086795

DI, Q., DAI, L., WANG, Y., ZANOBETTI, A., CHOIRAT, C., SCHWARTZ, J. D. and DOMINICI, F. (2017a). Association of short-term exposure to air pollution with mortality in older adults. *JAMA* **318** 2446–2456.

DI, Q., WANG, Y., ZANOBETTI, A., WANG, Y., KOUTRAKIS, P., CHOIRAT, C., DOMINICI, F. and SCHWARTZ, J. D. (2017b). Air pollution and mortality in the medicare population. *N. Engl. J. Med.* **376** 2513–2522.

FINLEY, A. O., DATTA, A., COOK, B. D., MORTON, D. C., ANDERSEN, H. E. and BANERJEE, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Statist.* **28** 401–414. MR3974889 https://doi.org/10.1080/10618600.2018.1537924

FRENCH, J. (2018). SpatialTools: Tools for spatial data analysis. R package version 1.0.4.

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

FULLER, W. A. (1987). *Measurement Error Models. Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0898653 https://doi.org/10.1002/9780470316665

HEFFERNAN, C., PENG, R., GENTNER, D. R., KOEHLER, K. and DATTA, A. (2023). Supplement to "A dynamic spatial filtering approach to mitigate underestimation bias in field calibrated low-cost sensor air pollution data." https://doi.org/10.1214/23-AOAS1751SUPP

HOLMES, J. and MORIARTY, W. (1999). Application of the generalized Pareto distribution to extreme value analysis in wind engineering. *J. Wind Eng. Ind. Aerodyn.* **83** 1–10.

JBAILY, A., ZHOU, X., LIU, J., LEE, T.-H., KAMAREDDINE, L., VERGUET, S. and DOMINICI, F. (2022). Air pollution exposure disparities across US population and income groups. *Nature* **601** 228–233.

JOHNSON, N. E., BONCZAK, B. and KONTOKOSTA, C. E. (2018). Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmos. Environ.* **184** 9–16.

JUNE, N., VAUGHAN, J., LEE, Y. and LAMB, B. K. (2021). Operational bias correction for PM2.5 using the AIRPACT air quality forecast system in the Pacific northwest. *J. Air Waste Manage. Assoc.* **71** 515–527.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D. J. Basic Engrg.* **82** 35–45. MR3931993

KIM, J., SHUSTERMAN, A. A., LIESCHKE, K. J., NEWMAN, C. and COHEN, R. C. (2018). The Berkeley atmospheric CO2 observation network: Field calibration and evaluation of low-cost air quality sensors. *Atmos. Meas. Tech.* **11** 1937–1946.

LARSON, T., HENDERSON, S. B. and BRAUER, M. (2009). Mobile monitoring of particle light absorption coefficient in an urban area as a basis for land use regression. *Environ. Sci. Technol.* **43** 4672–4678.

LEVY ZAMORA, M., XIONG, F., GENTNER, D., KERKEZ, B., KOHRMAN-GLASER, J. and KOEHLER, K. (2018). Field and laboratory evaluations of the low-cost Plantower particulate matter sensor. *Environ. Sci. Technol.* **53** 838–849.

LIM, C. C., KIM, H., VILCASSIM, M. R., THURSTON, G. D., GORDON, T., CHEN, L.-C., LEE, K., HEIMBINDER, M. and KIM, S.-Y. (2019). Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. *Environ. Int.* **131** 105022.

LU, Y. (2021). Beyond air pollution at home: Assessment of personal exposure to PM2.5 using activity-based travel demand model and low-cost air sensor network data. *Environ. Res.* **201** 111549.

MARDIA, K. V., GOODALL, C., REDFERN, E. J. and ALONSO, F. J. (1998). The Kriged Kalman filter. *TEST* **7** 217–285. MR1666999 https://doi.org/10.1007/BF02565111

MISKELL, G., SALMOND, J. A. and WILLIAMS, D. E. (2018). A solution to the problem of calibration of low-cost air quality measurement sensors in networks. *ACS Sensors* **3** 832–843.

NORDIO, F., KLOOG, I., COULL, B. A., CHUDNOVSKY, A., GRILLO, P., BERTAZZI, P. A., BACCARELLI, A. A. and SCHWARTZ, J. (2013). Estimating spatio-temporal resolved PM10 aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy. *Atmos. Environ.* **74** 227–236.

PETERS, T. M., NORRIS, G. A., VANDERPOOL, R. W., GEMMILL, D. B., WIENER, R. W., MURDOCH, R. W., MCELROY, F. F. and PITCHFORD, M. (2001). Field performance of PM2.5 federal reference method samplers. *Aerosol. Sci. Technol.* **34** 433–443.

PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131. MR0423667

ROMERO, Y., VELÁSQUEZ, R. M. A. and NOEL, J. (2020). Development of a multiple regression model to calibrate a low-cost sensor considering reference measurements and meteorological parameters. *Environ. Monitor. Assess.* **192** 1–11.

SAHU, S. K. and MARDIA, K. V. (2005). A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 223–244. MR2134608 https://doi.org/10.1111/j.1467-9876.2005.00480.x

SHI, L., STEENLAND, K., LI, H., LIU, P., ZHANG, Y., LYLES, R. H., REQUIA, W. J., ILANGO, S. D., CHANG, H. H. et al. (2021). A national cohort study (2000–2018) of long-term air pollution exposure and incident dementia in older adults in the United States. *Nat. Commun.* **12** 6754.

SI, M., XIONG, Y., DU, S. and DU, K. (2020). Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. *Atmos. Meas. Tech.* **13** 1693–1707.

TANG, X., ZHU, J., WANG, Z., WANG, M., GBAGUIDI, A., LI, J., SHAO, M., TANG, G. and JI, D. (2013). Inversion of CO emissions over Beijing and its surrounding areas with ensemble Kalman filter. *Atmos. Environ.* **81** 676–686.

TOPALOVIĆ, D. B., DAVIDOVIĆ, M. D., JOVANOVIĆ, M., BARTONOVA, A., RISTOVSKI, Z. and JOVAŠEVIĆ-STOJANOVIĆ, M. (2019). In search of an optimal in-field calibration method of low-cost gas sensors for ambient air pollutants: Comparison of linear, multilinear and artificial neural network approaches. *Atmos. Environ.* **213** 640–658.

TRYNER, J., L'ORANGE, C., MEHAFFY, J., MILLER-LIONBERG, D., HOFSTETTER, J. C., WILSON, A. and VOLCKENS, J. (2020). Laboratory evaluation of low-cost PurpleAir PM monitors and in-field correction using co-located portable filter samplers. *Atmos. Environ.* **220** 117067.

U.S. EPA (2019). Integrated Science Assessment (ISA) for Particulate Matter (Final Report). U.S. Environmental Protection Agency, Washington, DC.

U.S. EPA (2021). Air Data: Air Quality Data Collected at Outdoor Monitors Across the US. U.S. Environmental Protection Agency, Washington, DC. Available at https://www.epa.gov/outdoor-air-quality-data.

U.S. EPA (2022a). NAAQS Table. U.S. Environmental Protection Agency, Washington, DC. Available at https://www.epa.gov/criteria-air-pollutants/naaqs-table.

U.S. EPA (2022b). Policy Assessment for the Reconsideration of the National Ambient Air Quality Standards for Particulate Matter. U.S. Environmental Protection Agency, Washington, DC.

VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

VAN DER WAL, J. and JANSSEN, L. (2000). Analysis of spatial and temporal variations of PM10 concentrations in the Netherlands using Kalman filtering. *Atmos. Environ.* **34** 3675–3687.

WARD-CAVINESS, C. K., YAZDI, M. D., MOYER, J., WEAVER, A. M., CASCIO, W. E., DI, Q., SCHWARTZ, J. D. and DIAZ-SANCHEZ, D. (2021). Long-term exposure to particulate air pollution is associated with 30-day readmissions and hospital visits among patients with heart failure. *J. Amer. Heart Assoc.* **10** e019430. https://doi.org/10.1161/JAHA.120.019430

WEI, Y., WANG, Y., WU, X., DI, Q., SHI, L., KOUTRAKIS, P., ZANOBETTI, A., DOMINICI, F. and SCHWARTZ, J. D. (2020). Causal effects of air pollution on mortality rate in Massachusetts. *Amer. J. Epidemiol.* **189** 1316–1323.

WORLD HEALTH ORGANIZATION (2022). Ambient (outdoor) air pollution. World Health Organization. Available at https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

WU, H., TANG, X., WANG, Z., WU, L., LI, J., WANG, W., YANG, W. and ZHU, J. (2020). High-spatiotemporal-resolution inverse estimation of CO and NOx emission reductions during emission control periods with a modified ensemble Kalman filter. *Atmos. Environ.* **236** 117631.

WÜTHRICH, M., CIFUENTES, C. G., TRIMPE, S., MEIER, F., BOHG, J., ISSAC, J. and SCHAAL, S. (2016). Robust Gaussian filtering using a pseudo measurement. In 2016 *American Control Conference* (*ACC*) 3606–3613. IEEE, New York.

ZEGER, S. L., THOMAS, D., DOMINICI, F., SAMET, J. M., SCHWARTZ, J., DOCKERY, D. and COHEN, A. (2000). Exposure measurement error in time-series studies of air pollution: Concepts and consequences. *Environ. Health Perspect.* **108** 419–426.

ZHENG, T., BERGIN, M. H., SUTARIA, R., TRIPATHI, S. N., CALDOW, R. and CARLSON, D. E. (2019). Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in Delhi. *Atmos. Meas. Tech.* **12** 5161–5181.

ZIDEK, J. V., SHADDICK, G. and TAYLOR, C. G. (2014). Reducing estimation bias in adaptively changing monitoring networks with preferential site selection. *Ann. Appl. Stat.* **8** 1640–1670. MR3271347 https://doi.org/10.1214/14-AOAS745

ZIMMERMAN, D. L. and CRESSIE, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44** 27–43. MR1165570 https://doi.org/10.1007/BF00048668

ZIMMERMAN, D. L. and HOLLAND, D. M. (2005). Complementary co-Kriging: Spatial prediction using data combined from several environmental monitoring networks. *Environmetrics* **16** 219–234. MR2146909 https://doi.org/10.1002/env.699

ZIMMERMAN, N., PRESTO, A. A., KUMAR, S. P., GU, J., HAURYLIUK, A., ROBINSON, E. S., ROBINSON, A. L. and SUBRAMANIAN, R. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmos. Meas. Tech.* **11** 291–313.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3088–3111
https://doi.org/10.1214/23-AOAS1752

DATA-DRIVEN CHIMNEY FIRE RISK PREDICTION USING MACHINE LEARNING AND POINT PROCESS TOOLS

BY CHANGQING LU[1,a], MARIE-COLETTE VAN LIESHOUT[2,b], MAURITS DE GRAAF[3,c] AND PAUL VISSCHER[4,d]

[1]*Department of Applied Mathematics, University of Twente,* [a]*c.lu@utwente.nl*
[2]*Stochastics, Centrum Wiskunde & Informatica,* [b]*marie-colette.van.lieshout@cwi.nl*
[3]*Innovation Research & Technology, Thales Nederland B.V.,* [c]*maurits.de.graaf@cwi.nl*
[4]*Sector Strategy & Support, Brandweer Twente,* [d]*p.visscher@brandweertwente.nl*

Chimney fires constitute one of the most commonly occurring fire types. Precise prediction and prompt prevention are crucial in reducing the harm they cause. In this paper we develop a combined machine learning and statistical modelling process to predict fire risk. First, we use random forests and permutation importance techniques to identify the most informative explanatory variables. Second, we design a Poisson point process model and employ logistic regression estimation to estimate the parameters. Moreover, we validate the Poisson model assumption using second-order summary statistics and residuals. We implement the modelling process on data collected by the Twente Fire Brigade and obtain plausible predictions. Compared to similar studies, our approach has two advantages: (i) with random forests, we can select explanatory variables nonparametrically considering variable dependence; (ii) using logistic regression estimation, we can fit our statistical model efficiently by tuning it to focus on regions and times that are salient for fire risk.

REFERENCES

ALTMANN, A., TOLOŞI, L., SANDER, O. and LENGAUER, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics* **26** 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

BADDELEY, A., COEURJOLLY, J.-F., RUBAK, E. and WAAGEPETERSEN, R. (2014). Logistic regression for spatial Gibbs point processes. *Biometrika* **101** 377–392. MR3215354 https://doi.org/10.1093/biomet/ast060

BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns*: *Methodology and Applications with R*. CRC Press.

BADDELEY, A. and TURNER, R. (2000). Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Aust. N. Z. J. Stat.* **42** 283–322. MR1794056 https://doi.org/10.1111/1467-842X.00128

BADDELEY, A., TURNER, R., MØLLER, J. and HAZELTON, M. (2005). Residual analysis for spatial point processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 617–666. With discussion and a reply by the authors. MR2210685 https://doi.org/10.1111/j.1467-9868.2005.00519.x

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

BOUBETA, M., LOMBARDÍA, M. J., MAREY-PÉREZ, M. F. and MORALES, D. (2015). Prediction of forest fires occurrences with area-level Poisson mixed models. *J. Environ. Manag.* **154** 151–158. https://doi.org/10.1016/j.jenvman.2015.02.009

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

CHOIRUDDIN, A., COEURJOLLY, J.-F. and LETUÉ, F. (2018). Convex and non-convex regularization methods for spatial point processes intensity estimation. *Electron. J. Stat.* **12** 1210–1255. MR3780731 https://doi.org/10.1214/18-EJS1408

CHOIRUDDIN, A., COEURJOLLY, J.-F. and WAAGEPETERSEN, R. (2021). Information criteria for inhomogeneous spatial point processes. *Aust. N. Z. J. Stat.* **63** 119–143. MR4296145 https://doi.org/10.1111/anzs.12327

Key words and phrases. Fire prediction, K-function, logistic regression estimation, pair correlation function, Poisson point process, spatiotemporal point pattern, variable importance.

CLEVELAND, W. S., GROSSE, E. and SHYU, W. M. (1992). Local regression models. In *Statistical Models in S* 8, 1st ed. Wadsworth & Brooks/Cole.

COSTAFREDA-AUMEDES, S., COMAS, C. and VEGA-GARCIA, C. (2016). Spatio–temporal configurations of human-caused fires in Spain through point patterns. *Forests* **7** 185.

DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes*. *Volume II*, 2nd ed. Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

DEBEER, D. and STROBL, C. (2020). Conditional permutation importance revisited. *BMC Bioinform.* **21** 307. https://doi.org/10.1186/s12859-020-03622-2

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368. MR0773172 https://doi.org/10.1214/aos/1176346597

GABRIEL, E. and DIGGLE, P. J. (2009). Second-order analysis of inhomogeneous spatio–temporal point process data. *Stat. Neerl.* **63** 43–51. MR2656916 https://doi.org/10.1111/j.1467-9574.2008.00407.x

GODAMBE, V. P. and HEYDE, C. C. (1987). Quasi-likelihood and optimal estimation. *Int. Stat. Rev.* **55** 231–244. MR0963141 https://doi.org/10.2307/1403403

HERING, A. S., BELL, C. L. and GENTON, M. G. (2009). Modeling spatio–temporal wildfire ignition point patterns. *Environ. Ecol. Stat.* **16** 225–250. MR2668734 https://doi.org/10.1007/s10651-007-0080-6

HOTHORN, T., BUEHLMANN, P., DUDOIT, S., MOLINARO, A. and VAN DER LAAN, M. (2006). Survival ensembles. *Biostatistics* **7** 355–373.

HOTHORN, T., HORNIK, K. and ZEILEIS, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Statist.* **15** 651–674. MR2291267 https://doi.org/10.1198/106186006X133933

JAIN, P., COOGAN, S., SUBRAMANIAN, S., CROWLEY, M., TAYLOR, S. W. and FLANNIGAN, M. (2020). A review of machine learning applications in wildfire science and management. *Environ. Rev.* **28** 478–505.

JUAN VERDOY, P. (2021). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Comm. Statist. Simulation Comput.* **50** 2990–3030. MR4322119 https://doi.org/10.1080/03610918.2019.1618473

KOH, J., PIMONT, F., DUPUY, J.-L. and OPITZ, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *Ann. Appl. Stat.* **17** 560–582. MR4539044 https://doi.org/10.1214/22-aoas1642

LIESHOUT, M. N. M. VAN (2019). *Theory of Spatial Statistics*: *A Concise Introduction*. CRC Press, Boca Raton, FL.

LIESHOUT, M. N. M. VAN and LU, C. (2022). Infill asymptotics for logistic regression estimators for spatio–temporal point processes. arXiv:2208.12080.

LU, C., LIESHOUT, M. N. M. VAN, GRAAF, M. DE and VISSCHER, P. (2021). Chimney fire prediction based on explanatory environmental variables. In *The* 63*rd ISI World Statistics Congress* 288–291.

LU, C., LIESHOUT, M. N. M. VAN, GRAAF, M. DE and VISSCHER, P. (2023). Supplement to "Data-driven chimney fire risk prediction using machine learning and point process tools." https://doi.org/10.1214/23-AOAS1752SUPP

MALIK, A., RAO, M. R., PUPPALA, N., KOOURI, P., ANIL, V., THOTA, K., LIU, Q., CHIAO, S. and GAO, J. (2021). Data-driven wildfire risk prediction in northern California. *Atmosphere* **12** 109.

MCCULLAGH, P. and NELDER, J. A. (2019). *Generalized Linear Models*, 2nd ed. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MØLLER, J. and DÍAZ-AVALOS, C. (2010). Structured spatio–temporal shot-noise Cox point process models, with a view to modelling forest fires. *Scand. J. Stat.* **37** 2–25. MR2675937 https://doi.org/10.1111/j.1467-9469.2009.00670.x

NVBR (2010). *De brandweer over Morgen*. Nederlandse Vereniging voor Brandweer en Rampenbestrijding, Arnhem.

OHSER, J. and STOYAN, D. (1981). On the second-order and orientation analysis of planar stationary point processes. *Biom. J.* **23** 523–533. MR0635658 https://doi.org/10.1002/bimj.4710230602

PEREIRA, P., TURKMAN, K., TURKMAN, A., SÁ, A. and PEREIRA, J. (2013). Quantification of annual wildfire risk; a spatio–temporal point process approach. *Statistica* **73** 55–68.

PIMONT, F., FARGEON, H., OPITZ, T., RUFFAULT, J., BARBERO, R., MARTIN-STPAUL, N., RIGOLOT, E., RIVIÉRE, M. and DUPUY, J.-L. (2021). Prediction of regional wildfire activity in the probabilistic Bayesian framework of Firelihood. *Ecol. Appl.* **31** e02316. https://doi.org/10.1002/eap.2316

PREISLER, H., BRILLINGER, D., BURGAN, R. and BENOIT, J. (2004). Probability based models for estimation of wildfire risk. *Int. J. Wildland Fire* **13** 133–142.

RODRIGUES, M. and DE LA RIVA, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environ. Model. Softw.* **57** 192–201.

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

SAKR, G. E., ELHAJJ, I. H., MITRI, G. and WEJINYA, U. C. (2010). Artificial intelligence for forest fire prediction. In 2010 *IEEE/ASME International Conference on Advanced Intelligent Mechatronics* 1311–1316.

SATIR, O., BERBEROGLU, S. and DONMEZ, C. (2016). Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. *Geomatics, Natural Hazards and Risk* **7** 1645–1658.

SCHONLAU, M. and ZOU, R. (2020). The random forest algorithm for statistical learning. *Stata J.* **20** 3–29.

SCHOOL, M. L. (2018). A log-Gaussian Cox process for predicting chinmey fires at fire department Twente. Master's thesis, Univ. Twente.

SERRA, L., SAEZ, M., MATEU, J., VARGA, D., JUAN, P., DÍAZ-ÁVALOS, C. and RUE, H. (2014). Spatio–temporal log-Gaussian Cox processes for modelling wildfire occurrence: The case of Catalonia, 1994–2008. *Environ. Ecol. Stat.* **21** 531–563. MR3248538 https://doi.org/10.1007/s10651-013-0267-y

SILVAPULLE, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* **43** 310–313. MR0637943

STOJANOVA, D., KOBLER, A., OGRINC, P., ŽENKO, B. and DŽEROSKI, S. (2012). Estimating the risk of fire outbreaks in the natural environment. *Data Min. Knowl. Discov.* **24** 411–442.

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

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

STROBL, C., HOTHORN, T. and ZEILEIS, A. (2009). Party on! *R J.* **1** 14–17.

STROBL, C. and ZEILEIS, A. (2008). Danger: High power!—exploring the statistical properties of a test for random forest variable importance. In *COMPSTAT* 2008—*Proceedings in Computational Statistics* 59–66. Physica-Verlag/Springer, Heidelberg. MR2509600

THURMAN, A. L. and ZHU, J. (2014). Variable selection for spatial Poisson point processes via a regularization method. *Stat. Methodol.* **17** 113–125. MR3133589 https://doi.org/10.1016/j.stamet.2013.08.001

TURNER, R. (2009). Point pattern of forest fire locations. *Environ. Ecol. Stat.* **16** 197–223. MR2668733 https://doi.org/10.1007/s10651-007-0085-1

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York. https://doi.org/10.1007/978-0-387-21706-2

VER HOEF, J. M. (2012). Who invented the delta method? *Amer. Statist.* **66** 124–127. MR2968009 https://doi.org/10.1080/00031305.2012.687494

WONGVIBULSIN, S., WU, K. and ZEGER, S. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med. Res. Methodol.* **20**.

XU, H. and SCHOENBERG, F. P. (2011). Point process modeling of wildfire hazard in Los Angeles County, California. *Ann. Appl. Stat.* **5** 684–704. MR2840171 https://doi.org/10.1214/10-AOAS401

YANG, J., WEISBERG, P., DILTS, T., LOUDERMILK, L., SCHELLER, R., STANTON, A. and SKINNER, C. (2015). Predicting wildfire occurrence distribution with spatial point process models and its uncertainty assessment: A case study in the Lake Tahoe Basin, USA. *Int. J. Wildland Fire* **24** 390.

YE, R. (2011). Prediction of forest fires with Poisson models. *Can. J. For. Res.* **27** 1685–1694.

YUE, Y. and LOH, J. M. (2015). Variable selection for inhomogeneous spatial point process models. *Canad. J. Statist.* **43** 288–305. MR3353384 https://doi.org/10.1002/cjs.11244

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3112–3132
https://doi.org/10.1214/23-AOAS1753

WHEN ECOLOGICAL INDIVIDUAL HETEROGENEITY MODELS AND LARGE DATA COLLIDE: AN IMPORTANCE SAMPLING APPROACH

BY RUTH KING[1,a], BLANCA SARZO[2,c] AND VÍCTOR ELVIRA[1,b]

[1]*School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh,* [a]*Ruth.King@ed.ac.uk,* [b]*Victor.Elvira@ed.ac.uk*

[2]*Cavanilles Institute of Biodiversity and Evolutive Biology, Department of Microbiology and Ecology, University of Valencia,* [c]*Blanca.Sarzo@uv.es*

We consider the challenges that arise when fitting ecological individual heterogeneity models to "large" data sets. In particular, we focus on (continuous-valued) random effect models commonly used to describe individual heterogeneity present in ecological populations within the context of capture–recapture data, although the approach is more widely applicable to more general latent variable models. Within such models the associated likelihood is expressible only as an analytically intractable integral. Common techniques for fitting such models to data include, for example, the use of numerical approximations for the integral or a Bayesian data augmentation approach. However, as the size of the data set increases (i.e., the number of individuals increases), these computational tools may become computationally infeasible. We present an efficient Bayesian model-fitting approach, whereby we initially sample from the posterior distribution of a smaller subsample of the data, before correcting this sample to obtain estimates of the posterior distribution of the full data set using an importance sampling approach. We consider several practical issues, including the subsampling mechanism, computational efficiencies (including the ability to parallelise the algorithm) and combining subsampling estimates using multiple subsampled data sets. We initially demonstrate the feasibility (and accuracy) of the approach via simulated data before considering a challenging real data set of approximately 30,000 guillemots and, using the proposed algorithm, obtain posterior estimates of the model parameters in substantially reduced computational time, compared to the standard Bayesian model-fitting approach.

REFERENCES

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

ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648 https://doi.org/10.1214/07-AOS574

BARDENET, R., DOUCET, A. and HOLMES, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **18** Paper No. 47, 43. MR3670492

BROOKS, S., GELMAN, A., JONES, G. L. and MENG, X.-L., eds. (2011). *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR2742422 https://doi.org/10.1201/b10905

BUTLER, J. and MOFFIT, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica* 761–764.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BENTANCOURT, M., BRUBAKER, M., GUO, J. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76**.

COULL, B. A. and AGRESTI, A. (1999). The use of mixed logit models to reflect heterogeneity in capture–recapture studies. *Biometrics* **55** 294–301. https://doi.org/10.1111/j.0006-341x.1999.00294.x

DE VALPINE, P. (2002). Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state–space models. *Bull. Mar. Sci.* **70** 455–471.

Key words and phrases. Capture–recapture, Cormack–Jolly–Seber model, importance sampling, individual heterogeneity, intractable likelihood, random effects.

DE VALPINE, P. (2004). Monte Carlo state–space likelihoods by weighted posterior kernel density estimation. *J. Amer. Statist. Assoc.* **99** 523–536. MR2062837 https://doi.org/10.1198/016214504000000476

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

DOUC, R., GUILLIN, A., MARIN, J.-M. and ROBERT, C. P. (2007). Minimum variance importance sampling via population Monte Carlo. *ESAIM Probab. Stat.* **11** 427–447. MR2339302 https://doi.org/10.1051/ps:2007028

ELVIRA, V. and MARTINO, L. (2021). Advances in importance sampling. *Wiley StatsRef: Statistics Reference Online* 1–14.

ELVIRA, V., MARTINO, L. and CLOSAS, P. (2021). Importance Gaussian quadrature. *IEEE Trans. Signal Process.* **69** 474–488. MR4213358 https://doi.org/10.1109/TSP.2020.3045526

FRANCIS, C. M. and SAUROLA, P. (2009). Estimating demographic parameters from complex data sets: A comparison of Bayesian hierarchical and maximum-likelihood methods for estimating survival probabilities of tawny owls, *Strix aluco* in Finland. In *Modeling Demographic Processes in Marked Populations* (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.) 617–637. Springer, Boston, MA.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**. Chapman & Hall/CRC, Boca Raton, FL, USA. MR2027492

GIMENEZ, O., BONNER, S., KING, R., PARKER, R. A., BROOKS, S. P., JAMIESON, L. E., GROSBOIS, V., MORGAN, B. J. T. and THOMAS, L. (2009). WinBUGS for population ecologists: Bayesian modelling using Markov chain Monte Carlo (MCMC) methods. In *Modeling Demographic Processes in Marked Populations* (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.) 885–918. Springer, Boston, MA.

GIMENEZ, O., CAM, E. and GAILLARD, J.-M. (2017). Individual heterogeneity and capture–recapture models: What, why and how? *Oikos* **127** 664–686.

GIMENEZ, O. and CHOQUET, R. (2010). Individual heterogeneity in studies on marked animals using numerical integration: Capture–recapture mixed models. *Ecology* **91** 951–957. https://doi.org/10.1890/09-1903.1

HANKIN, D., MOHR, M. and NEWMAN, K. (2019). *Sampling Theory*. Oxford Univ. Press.

HEDEKER, D. and GIBBONS, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** 933–944.

HERLIANSYAH, R., KING, R. and KING, S. (2022). Laplace approximations for capture–recapture models in the presence of individual heterogeneity. *J. Agric. Biol. Environ. Stat.* **27** 401–418. MR4458685 https://doi.org/10.1007/s13253-022-00486-2

HESTBECK, J. B., NICHOLS, J. D. and MALECKI, R. A. (1991). Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. *Ecology* **72** 523–533. https://doi.org/10.2307/2937193

HUGGINS, J., CAMPBELL, T. and BRODERICK, T. (2016). Coresets for scalable Bayesian logistic regression. *Adv. Neural Inf. Process. Syst.* **29**.

IONIDES, E. L. (2008). Truncated importance sampling. *J. Comput. Graph. Statist.* **17** 295–311. MR2439961 https://doi.org/10.1198/106186008X320456

KÉRY, M. and SCHAUB, M. (2011). *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press.

KING, R. (2014). Statistical ecology. *Annu. Rev. Stat. Appl.* **1** 401–426.

KING, R. and BROOKS, S. P. (2008). On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* **64** 816–824. MR2526632 https://doi.org/10.1111/j.1541-0420.2007.00938.x

KING, R., MCCLINTOCK, B. T., KIDNEY, D. and BORCHERS, D. (2016). Capture–recapture abundance estimation using a semi-complete data likelihood approach. *Ann. Appl. Stat.* **10** 264–285. MR3480496 https://doi.org/10.1214/15-AOAS890

KING, R., MORGAN, B. J. T., GIMÉNEZ, O. and BROOKS, S. P. (2010). *Bayesian Analysis for Population Ecology*. CRC Press.

KING, R., SARZO, B. and ELVIRA, V. (2023). Supplement to "When ecological individual heterogeneity models and large data collide: An importance sampling approach." https://doi.org/10.1214/23-AOAS1753SUPPA, https://doi.org/10.1214/23-AOAS1753SUPPB

LIU, Q. and PIERCE, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81** 624–629. MR1311107 https://doi.org/10.1093/biomet/81.3.624

LUENGO, D., MARTINO, L., ELVIRA, V. and BUGALLO, M. (2018). Efficient linear fusion of partial estimators. *Digit. Signal Process.* **78** 265–283.

LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10** 325–337.

MCCREA, R. S. and MORGAN, B. J. T. (2015). *Analysis of Capture–Recapture Data. Chapman & Hall/CRC Interdisciplinary Statistics Series*. CRC Press, Boca Raton, FL. MR3330977

NGUYEN, T. L. T., SEPTIER, F., PETERS, G. W. and DELIGNON, Y. (2014). Improving SMC sampler estimate by recycling all past simulated particles. In *Statistical Signal Processing* (*SSP*), 2014 *IEEE Workshop on* 117–120. IEEE.

OLSSON, O. and HENTATI-SUNDBERG, J. (2017). Population trends and status of four seabird species (*Uria aalge*, *Alca torda*, *Larus fuscus*, *Larus argentatus*) at Stora Karlsö in the Baltic Sea. *Ornys Svecica* **27** 64–93.

OWEN, A. (2013). Monte Carlo theory, methods and examples. http://statweb.stanford.edu/~owen/mc/.

PLEDGER, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics* **56** 434–442. https://doi.org/10.1111/j.0006-341x.2000.00434.x

PLEDGER, S., POLLOCK, K. H. and NORRIS, J. L. (2003). Open capture–recapture models with heterogeneity. I. Cormack–Jolly–Seber model. *Biometrics* **59** 786–794. MR2025102 https://doi.org/10.1111/j.0006-341X.2003.00092.x

PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the* 3*rd International Workshop on Distributed Statistical Computing* **124** 1–9.

ROBERT, C. P., ELVIRA, V., TAWN, N. and WU, C. (2018). Accelerating MCMC algorithms. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **10** e1435, 14. MR3850448 https://doi.org/10.1002/wics.1435

ROYLE, J. A. (2008). Modeling individual effects in the Cormack–Jolly–Seber model: A state–space formulation. *Biometrics* **64** 364–370, 664. MR2432405 https://doi.org/10.1111/j.1541-0420.2007.00891.x

SARZO, B., ARMERO, C., CONESA, D., HENTATI-SUNDBERG, J. and OLSSON, O. (2019). Bayesian immature survival analysis of the largest colony of Common murre *Uria aalge* in the Baltic sea. *Waterbirds* **42** 304–313.

SARZO, B., KING, R., CONESA, D. and HENTATI-SUNDBERG, J. (2021). Correcting bias in survival probabilities for partially monitored populations via integrated models. *J. Agric. Biol. Environ. Stat.* **26** 200–219. MR4257013 https://doi.org/10.1007/s13253-020-00423-1

SEBER, G. A. F. and SCHOFIELD, M. R. (2019). *Capture–Recapture*: *Parameter Estimation for Open Animal Populations. Statistics for Biology and Health*. Springer, Cham. MR3967717 https://doi.org/10.1007/978-3-030-18187-1

TOKDAR, S. T. and KASS, R. E. (2010). Importance sampling: A review. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **2** 54–60.

TRAN, M.-N., SCHARTH, M., PITT, M. K. and KOHN, R. (2016). Importance sampling squared for Bayesian inference in latent variable models. arXiv:1309.3339.

VAN DE SCHOOT, R., DEPAOLI, S., KING, R., KRAMER, B., MÄRTENS, K., TADESSE, M. G., VANNUCCI, M., GELMAN, A., VEEN, D. et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers* **1** 1–26.

VEHTARI, A., SIMPSON, D., GELMAN, A., YAO, Y. and GABRY, J. (2015). Pareto smoothed importance sampling. arXiv:1507.02646.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3133–3152
https://doi.org/10.1214/23-AOAS1754

DESIGN-BASED MAPPING OF LAND USE/LAND COVER CLASSES WITH BOOTSTRAP ESTIMATION OF PRECISION BY NEAREST-NEIGHBOUR INTERPOLATION

BY AGNESE MARCELLI[1,a], ROSA MARIA DI BIASE[2,b], PIERMARIA CORONA[3,c], STEPHEN V. STEHMAN[4,d] AND LORENZO FATTORINI[5,e]

[1]*Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia,* [a]*agnese.marcelli@unitus.it*
[2]*Department of Sociology and Social Research, University of Milano Bicocca,* [b]*rosamaria.dibiase@unimib.it*
[3]*CREA, Research Centre for Forestry and Wood,* [c]*piermaria.corona@crea.gov.it*
[4]*Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry,* [d]*svstehma@syr.edu*
[5]*Department of Economic and Statistics, University of Siena,* [e]*lorenzo.fattorini@unisi.it*

Land use/land cover mapping is usually performed by classifying satellite imagery (e.g., Landsat, Sentinel) for the whole survey region using classification algorithms implemented with training data. Subsequently, probabilistic samples are usually implemented with the main purpose of assessing the accuracy of these maps by comparing the map class and the ground condition determined for the sampled units. The main proposal of this paper is to directly exploit these probabilistic samples to estimate the land use/land cover class at any location of the survey region in a design-based framework by the well-known nearest-neighbour interpolator. For the first time, the design-based consistency of nearest-neighbour maps (i.e., categorical variables) is theoretically proven and a pseudo-population bootstrap estimator of their precision is proposed and discussed. These nearest-neighbour maps provide the ability to place mapping within a rigorous design-based inference framework, in contrast to most traditional mapping approaches which often are implemented with no inferential basis or by necessity (due to lack of a probabilistic sample) model-based inference. A simulation study is performed on an estimated land use map in Southern Tuscany (Italy)—taken as the true map—to check the finite-sample performance of the proposal as well as the matching of the area coverage estimates arising from the map with those achieved by traditional estimators. The Italian land use map arising from the IUTI surveys and the U.S. land cover map arising from the LCMAP program are considered as case studies.

REFERENCES

AL-DOSKI, J., MANSOR, S. B., SAN, H. P. and KHUZAIMAH, Z. (2020). Land cover mapping using remote sensing data. *Am. J. Geogr. Inf. Syst.* **9** 33–45.

AUCH, R. F., WELLINGTON, D. F., TAYLOR, J. L., STEHMAN, S. V., TOLLERUD, H. J., BROWN, J. F., LOVELAND, T. R., PENGRA, B. W., HORTON, J. A. et al. (2022). Conterminous United States land-cover change (1985–2016): New insights from annual time series. *Land* **11** 298.

BARABESI, L. (2003). A Monte Carlo integration approach to Horvitz–Thompson estimation in replicated environmental designs. *Metron* **LXI** 355–374.

BARABESI, L., FRANCESCHI, S. and MARCHESELLI, M. (2012). Properties of design-based estimation under stratified spatial sampling with application to canopy coverage estimation. *Ann. Appl. Stat.* **6** 210–228.

BROWN, J. F., TOLLERUD, H. J., BARBER, C. P., ZHOU, Q., DWYER, J. L., VOGELMAN, J. E., LOVELAND, T. R., WOODCOK, C. E., STEHMAN, S. V. et al. (2020). Lessons learned implementing an operational continuous United States national land change monitoring capability: The land change monitoring, assessment, and projection (LCMAP) approach. *Remote Sens. Environ.* **238** 111356.

Key words and phrases. Spatial sampling, spatial interpolation, consistency, pseudo-population bootstrap, simulation, case studies.

CIHLAR, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities. *Int. J. Remote Sens.* **21** 1093–1114.

COMBER, A., SEE, L., FRITZ, S., VAN DER VELDE, M., PERGER, C. and FOODY, G. M. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *Int. J. Appl. Earth Obs. Geoinf.* **23** 37–48.

CORONA, P., BARBATI, A., TOMAO, A., BERTANI, R., VALENTINI, R., MARCHETTI, M., FATTORINI, L. and PERUGINI, L. (2012). Land use inventory as framework for environmental accounting: An application in Italy. *iForest* **5** 204–209.

DI BIASE, R. M., FATTORINI, L., FRANCESCHI, S., GROTTI, M., PULETTI, N. and CORONA, P. (2022). From model selection to maps: A completely design-based data-driven inference for mapping forest resources. *Environmetrics* **33** e2750.

FATTORINI, L. (2015). Design-based methodological advances to support national forest inventories: A review of recent proposals. *iForest* **8** 6–11.

FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2004). Two-phase estimation of coverages with second-phase corrections. *Environmetrics* **15** 357–368.

FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2006). A three-phase sampling strategy for large-scale multiresource forest inventories. *J. Agric. Biol. Environ. Stat.* **11** 296–316.

FATTORINI, L., MARCHESELLI, M., PISANI, C. and PRATELLI, L. (2022). Design-based properties of the nearest neighbor spatial interpolator and its bootstrap mean squared error estimator. *Biometrics* **78** 1454–1463.

FITZPATRICK-LINS, K. (1981). Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogramm. Eng. Remote Sens.* **47** 343–351.

GRAFSTRÖM, A., SAARELA, S. and ENE, L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Can. J. For. Res.* **44** 1156–1164.

GREGOIRE, T. G. and VALENTINE, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. CRC Press/CRC, Boca Raton, FL.

HANSEN, M., DUBAYAH, R. and DE FRIES, R. (1996). Classification trees: An alternative to traditional land cover classifier. *Int. J. Remote Sens.* **17** 1075–1081.

INTERNATIONAL PANEL ON CLIMATE CHANGE (2003). *Good Practice Guidance for Land Use, Land Use Change and Forestry. IPCC National Greenhouse Gas Inventories Program.*

ISPRA (2014). Italian Greenhouse Gas Inventory 1990–2012. National Inventory Report 2014 ISPRA Rapporti 198/14.

KHATAMI, R., MOUNTRAKIS, G. and STEHMAN, S. V. (2017). Mapping per-pixel predicted accuracy of classified remote sensing images. *Remote Sens. Environ.* **191** 156–167.

MANNAN, B., ROY, J. and RAY, A. K. (1998). Fuzzy ARTMAP supervised classification of multi-spectral remotely-sensed images. *Int. J. Remote Sens.* **19** 767–774.

MARCELLI, A., FATTORINI, L. and FRANCESCHI, S. (2022). Harmonization of design-based mapping for spatial populations. *Stoch. Environ. Res. Risk Assess.*. **36** 3171–3182.

MARCELLI, A., DI BIASE, R. M., CORONA, P., STEHMAN, S. V. and FATTORINI, L. (2023). Supplement to "Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation." https://doi.org/10.1214/23-AOAS1754SUPP

MCROBERTS, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sens. Environ.* **115** 715–724.

NATIONAL RESEARCH COUNCIL (2001). *Grand Challenges in Environmental Sciences*. The National Academy Press, Washington, DC.

NGUYEN, H. T. T., DOAN, T. M., TOMPPO, E. and MCROBERTS, R. E. (2020). Land use/land cover mapping using multitemporal Sentinel-2 imagery and four classification methods—a case study from Dak Nong, Vietnam. *Remote Sens.* **12** 1367.

NUSSER, S. M. and KLAAS, E. E. (2003). Survey methods for assessing land cover map accuracy. *Environ. Ecol. Stat.* **10** 309–331.

OLOFSSON, P., FOODY, G. M., HEROLD, M., STEHMAN, S. V., WOODCOCK, C. E. and WULDER, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **148** 42–57.

OPSOMER, J. D., BREIDT, F. J., MOISEN, G. G. and KAUERMANN, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *J. Amer. Statist. Assoc.* **102** 400–409.

PENGRA, B. W., STEHMAN, S. V., HORTON, J. A., DOCKTER, D. J., SCHROEDER, T. A., YANG, Z., COHEN, W. B., HEALEY, S. P. and LOVELAND, T. R. (2020). Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. *Remote Sens. Environ.* **238** 111261.

QUATEMBER, A. (2015). *Pseudo-Populations. A Basic Concept in Statistical Surveys*. Springer, Cham.

RIZZO, M. and GASPARINI, P. (2022). Land use and land cover photointerpretation. In *Italian National Forest Inventory-Methods and Results of the Third Survey* (P. Gasparini, L. Di Cosmo, A. Floris and D. De Laurentis, eds.) 49–59. Springer, Cham, CH.

RODRÍGUEZ-JEANGROS, N., HERING, A. S., KAISER, T. and MCCRAY, J. (2016). Fusing multiple existing space-time land cover products. *Environmetrics* **28** e2429.

RODRÍGUEZ-JEANGROS, N., HERING, A. S., KAISER, T. and MCCRAY, J. (2017). SCaMF-RM: A fused high-resolution land cover product of the Rocky Mountains. *Remote Sens.* **9** 1015.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. *Springer Series in Statistics*. Springer, New York.

STEHMAN, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **62** 77–89.

STEHMAN, S. V. (2009). Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **30** 5243–5272.

STEHMAN, S. V. and CZAPLEWSKI, R. L. (1998). Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sens. Environ.* **64** 331–344.

STEHMAN, S. V. and FOODY, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sens. Environ.* **231** 111199.

STEHMAN, S. V., PENGRA, B. P., HORTON, J. A. and WELLINGTON, D. F. (2021). Validation of the United States geological survey's land change monitoring, assessment and projection (LCMAP) annual land cover products 1985–2017. *Remote Sens. Environ.* **265** 112646.

TOMPPO, L. M., GSCHWANTNER, T. and MCROBERTS, R. E. (2010). *National Forest Inventories*: *Pathways for Common Reporting*. Springer, Heidelberg, DE.

TURNER, B. L., LAMBIM, E. F. and REENBERG, A. (2007). The emergence of land change science for global environmental change and sustainability. *Proc. Natl. Acad. Sci. USA* **104** 20666–20671.

VAN DER MEER, F. (1995). Spectral unmixing of landsat thematic mapper data. *Int. J. Remote Sens.* **16** 3189–3194.

YOOL, S. R. (1998). Land cover classification in rugged areas using simulated moderate-resolution remote sensor data and an artificial neural network. *Int. J. Remote Sens.* **19** 85–96.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3153–3172
https://doi.org/10.1214/23-AOAS1755

IDENTIFYING BOUNDARIES IN SPATIALLY CONTINUOUS RISK SURFACES FROM SPATIALLY AGGREGATED DISEASE COUNT DATA

BY DUNCAN LEE[a]

School of Mathematics and Statistics, University of Glasgow, [a]*Duncan.Lee@glasgow.ac.uk*

Spatially aggregated disease-count data relating to a set of nonoverlapping areal units are often used to make inference on population-level disease risk. This includes the identification of risk boundaries, which are locations where there is a sizeable change in risk between geographically neighbouring areal units. Existing studies provide spatially discrete inference on the areal unit footprint, which forces the boundaries to coincide with the entire geographical border between neighbouring units. This paper is the first to relax these assumptions by estimating disease risk and the locations of risk boundaries on a grid of square pixels covering the study region that can be made arbitrarily small to approximate a spatially continuous surface. We propose a two-stage approach that first fits a Bayesian spatiotemporal realignment model to estimate disease risk at the grid level and then identifies boundaries in this surface using edge detection algorithms from computer vision. This novel methodological fusion is motivated by a new study of respiratory hospitalisation risk in Glasgow, Scotland, between 2008 and 2017, and we identify numerous risk boundaries across the city.

REFERENCES

BANERJEE, S. and GELFAND, A. E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *J. Amer. Statist. Assoc.* **101** 1487–1501. MR2279474 https://doi.org/10.1198/016214506000000041

BERCHUK, S. I., MWANZA, J.-C. and WARREN, J. L. (2019). Diagnosing glaucoma progression with visual field data using a spatiotemporal boundary detection method. *J. Amer. Statist. Assoc.* **114** 1063–1074. MR4011758 https://doi.org/10.1080/01621459.2018.1537911

BERNARDINELLI, L., CLAYTON, D., PASCUTTO, C., MONTOMOLI, C., GHISLANDI, M. and SONGINI, M. (1995). Bayesian analysis of space-time variation in disease risk. *Stat. Med.* **14** 2433–2443. https://doi.org/10.1002/sim.4780142112

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

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

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

CANNY, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8** 679–698.

DEAN, N., DONG, G., PIEKUT, A. and PRYCE, G. (2019). Frontiers in residential segregation: Understanding neighbourhood boundaries and their impacts. *Tijdschrift voor Economische en Sociale Geografie* **110** 271–288.

FISHER, T. J., ZHANG, J., COLEGATE, S. P. and VANNI, M. J. (2022). Detecting and modeling changes in a time series of proportions. *Ann. Appl. Stat.* **16** 477–494. MR4400519 https://doi.org/10.1214/21-aoas1509

FLOWERDEW, R. and GREEN, M. (1989). Statistical methods for inference between incompatible zonal systems. In *Accuracy of Spatial Databases* 239–247. Taylor & Francis, London.

FLOWERDEW, R. and GREEN, M. (1993). Developments in areal interpolation methods and GIS. In *Geographic Information Systems*, *Spatial Modelling and Policy Evaluation* 73–84. Springer, Berlin.

Key words and phrases. Bayesian inference, disease-risk modelling, edge-detection algorithms, spatially continuous inference.

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

GRAMATICA, M., CONGDON, P. and LIVERANI, S. (2021). Bayesian modelling for spatially misaligned health areal data: A multiple membership approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 645–666. MR4275840 https://doi.org/10.1111/rssc.12480

KNORR-HELD, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Stat. Med.* **19** 2555–2567.

LEE, D. (2023). Supplement to "Identifying boundaries in spatially continuous risk surfaces from spatially aggregated disease count data." https://doi.org/10.1214/23-AOAS1755SUPPA, https://doi.org/10.1214/23-AOAS1755SUPPB

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

LEE, D. and MITCHELL, R. (2012). Boundary detection in disease mapping studies. *Biostatistics* **13** 415–426. https://doi.org/10.1093/biostatistics/kxr036

LEE, D. and MITCHELL, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 593–608. MR3083913 https://doi.org/10.1111/rssc.12009

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

LU, H. and CARLIN, B. (2005). Bayesian areal Wombling for geographical boundary analysis. *Geogr. Anal.* **37** 265–285.

MA, H., CARLIN, B. P. and BANERJEE, S. (2010). Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis. *Biometrics* **66** 355–364. MR2758815 https://doi.org/10.1111/j.1541-0420.2009.01291.x

MACKENBACH, J., VALVERDE, J., ARTNIK, B. et al. (2018). Trends in health inequalities in 27 European countries. *Proc. Natl. Acad. Sci. USA* **115** 6440–6445.

MARR, D. and HILDRETH, E. (1980). Theory of edge detection. *Proc. R. Soc. Lond., B Biol. Sci.* **207** 187–217. https://doi.org/10.1098/rspb.1980.0020

MITCHELL, R. and LEE, D. (2014). Is there really a 'wrong side of the tracks' in urban areas and does it matter for spatial analysis? *Ann. Assoc. Amer. Geogr.* **104** 432–443.

MUGGLIN, A. S. and CARLIN, B. P. (1998). Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *J. Agric. Biol. Environ. Stat.* **3** 111–130. MR1816411 https://doi.org/10.2307/1400646

MUNTARINA, K., SHORIF, S. and UDDIN, M. (2022). Notes on edge detection approaches. *Evolving Systems* **13** 169–182.

NHS HEALTH SCOTLAND (2016). Health inequalities—what are they and how do we reduce them? Available at http://www.healthscotland.scot/media/1086/health-inequalities-what-are-they-how-do-we-reduce-them-mar16.pdf.

PREWITT, J. (1970). Object enhancement and extraction. In *Picture Processing and Psychopictorics* (B. Lipkin and A. Rosenfeld, eds.) 75–149. Academic Press, New York.

QU, K., BRADLEY, J. R. and NIU, X. (2021). Boundary detection using a Bayesian hierarchical model for multiscale spatial data. *Technometrics* **63** 64–76. MR4205692 https://doi.org/10.1080/00401706.2019.1677268

RUSHWORTH, A., LEE, D. and MITCHELL, R. (2014). A spatiotemporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatiotemporal Epidemiology* **10** 29–38.

RUSHWORTH, A., LEE, D. and SARRAN, C. (2017). An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 141–157. MR3611681 https://doi.org/10.1111/rssc.12155

SMITH, S. and BRADY, M. (1997). SUSAN—a new approach to low level image processing. *Int. J. Comput. Vis.* **23** 45–78.

SOBEL, I. and FELDMAN, G. (1968). A 3×3 isotropic gradient operator for image processing. Presentation at Stanford A.I. project 1968.

SYRING, N. and LI, M. (2017). BayesBD: An R package for Bayesian inference on image boundaries. *R J.* **9** 149–162.

TAYLOR, B. M., ANDRADE-PACHECO, R. and STURROCK, H. J. W. (2018). Continuous inference for aggregated point process data. *J. Roy. Statist. Soc. Ser. A* **181** 1125–1150. MR3876385 https://doi.org/10.1111/rssa.12347

WAKEFIELD, J. and KIM, A. (2013). A Bayesian model for cluster detection. *Biostatistics* **14** 752–765.

WALLER, L., CARLIN, B., XIA, H. and GELFAND, E. (1997). Hierarchical spatiotemporal mapping of disease rates. *J. Amer. Statist. Assoc.* **92** 607–617.

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

ZHIDING, Y., CHEN, F., MING-YU, L. and SRIKUMAR, R. (2017). CASENet: Deep category-aware semantic edge detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 5964–5973.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3173–3194
https://doi.org/10.1214/23-AOAS1756

STOCHASTIC DECLUSTERING OF EARTHQUAKES WITH THE SPATIOTEMPORAL RENEWAL ETAS MODEL

BY TOM STINDL[a] AND FENG CHEN[b]

Department of Statistics, UNSW Sydney, [a]*t.stindl@unsw.edu.au,* [b]*feng.chen@unsw.edu.au*

Modeling and forecasting earthquakes is challenging due to the complex interplay and clustering of main-shocks and aftershocks. The epidemic-type aftershock sequence (ETAS) model represents the conditional intensity of earthquakes as the superposition of a background and aftershock rate which allows for the declustering of the earthquakes. Its success has led to the development of numerous versions of the ETAS model. Among these extensions is the renewal ETAS (RETAS) model, which has shown promising potential. The RETAS model endows the main-shock arrival process with a renewal process, which serves as an alternative to the homogeneous Poisson process. Model fitting is performed using likelihood-based estimation by directly optimizing the exact likelihood. However, inferring the branching structure from the fitted RETAS model remains a challenging task since the declustering algorithm that is currently available for the ETAS model is not directly applicable. Therefore, this article develops an iterative algorithm to calculate the smoothed main- and aftershock probabilities, conditional on all available information contained in the catalog. Consequently, an estimate of the background spatial intensity function and model parameters can be obtained using an iterative semiparametric procedure with the smoothing parameters selected using information criteria. The methods proposed herein are illustrated on simulated data and a New Zealand earthquake catalog.

REFERENCES

ADELFIO, G. and CHIODI, M. (2021). Including covariates in a space-time point process with application to seismicity. *Stat. Methods Appl.* **30** 947–971. MR4308371 https://doi.org/10.1007/s10260-020-00543-5

AKAIKE, H. (1971). Autoregressive model fitting for control. *Ann. Inst. Statist. Math.* **23** 163–180. MR0348947 https://doi.org/10.1007/BF02479221

BURNHAM, K. P. and ANDERSON, D. R. (2002). A practical information-theoretic approach. *Model Selection and Multimodel Inference* **2** 70–71.

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

CHENG, Y., DUNDAR, M. and MOHLER, G. (2018). A coupled ETAS-I^2GMM point process with applications to seismic fault detection. *Ann. Appl. Stat.* **12** 1853–1870. MR3852700 https://doi.org/10.1214/18-AOAS1134

CHIODI, M. and ADELFIO, G. (2017). Mixed non-parametric and parametric estimation techniques in R package etasFLP for earthquakes' description. *J. Stat. Softw.* **76** 1–29.

CLARK, N. J. and DIXON, P. M. (2018). Modeling and estimation for self-exciting spatio-temporal models of terrorist activity. *Ann. Appl. Stat.* **12** 633–653. MR3773408 https://doi.org/10.1214/17-AOAS1112

CONSOLE, R., MURRU, M. and LOMBARDI, A. M. (2003). Refining earthquake clustering models. *J. Geophys. Res., Solid Earth* **108**.

FOX, E. W., SCHOENBERG, F. P. and GORDON, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.* **10** 1725–1756. MR3553242 https://doi.org/10.1214/16-AOAS957

FOX, E. W., SHORT, M. B., SCHOENBERG, F. P., CORONGES, K. D. and BERTOZZI, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Amer. Statist. Assoc.* **111** 564–584. MR3538687 https://doi.org/10.1080/01621459.2015.1135802

GENEST, C., QUESSY, J.-F. and REMILLARD, B. (2007). Asymptotic local efficiency of Cramér-von Mises tests for multivariate independence. *Ann. Statist.* **35** 166–191. MR2332273 https://doi.org/10.1214/009053606000000984

Key words and phrases. Point process, renewal process, semiparametric, self-exciting, seismology.

GENEST, C. and RÉMILLARD, B. (2004). Tests of independence and randomness based on the empirical copula process. *TEST* **13** 335–370. MR2154005 https://doi.org/10.1007/BF02595777
GUO, Y., ZHUANG, J. and ZHOU, S. (2015). An improved space-time ETAS model for inverting the rupture geometry from seismicity triggering. *J. Geophys. Res., Solid Earth* **120** 3309–3323.
GUTENBERG, B. and RICHTER, C. F. (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Amer.* **34** 185–188.
HARTE, D. S. (2013). Bias in fitting the ETAS model: A case study based on New Zealand seismicity. *Geophys. J. Int.* **192** 390–412.
HARTE, D. S. (2014). An ETAS model with varying productivity rates. *Geophys. J. Int.* **198** 270–284.
HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. MR1016020 https://doi.org/10.1093/biomet/76.2.297
MCCLOUD, N. and PARMETER, C. F. (2020). Determining the number of effective parameters in kernel density estimation. *Comput. Statist. Data Anal.* **143** 106843. MR4016964 https://doi.org/10.1016/j.csda.2019.106843
MEYER, S., ELIAS, J. and HÖHLE, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68** 607–616. MR2959628 https://doi.org/10.1111/j.1541-0420.2011.01684.x
MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.
MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546
MUSMECI, F. and VERE-JONES, D. (1992). A space-time clustering model for historical earthquakes. *Ann. Inst. Statist. Math.* **44** 1–11.
OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.
OGATA, Y. (1998). Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.
OGATA, Y. (2011). Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity. *Earth Planets Space* **63** 6.
OMORI, F. (1894). On the aftershocks of earthquakes. *J. Coll. Sci., Imp. Univ. Tokyo* **7** 111–120.
PENG, R. D., SCHOENBERG, F. P. and WOODS, J. A. (2005). A space-time conditional intensity model for evaluating a wildfire hazard index. *J. Amer. Statist. Assoc.* **100** 26–35. MR2166067 https://doi.org/10.1198/016214504000001763
REID, H. F. (1910). The California earthquake of April 18, 1906. Volume II. The Mechanics of the Earthquake. Washington DC: Carnegie Institution of Washington, Publication No. 87.
SCHOENBERG, F. P., HOFFMANN, M. and HARRIGAN, R. J. (2019). A recursive point process model for infectious diseases. *Ann. Inst. Statist. Math.* **71** 1271–1287. MR3993533 https://doi.org/10.1007/s10463-018-0690-9
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
STINDL, T. and CHEN, F. (2022). Spatiotemporal ETAS model with a renewal main-shock arrival process. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 1356–1380. MR4511114
STINDL, T. and CHEN, F. (2023). Supplement to "Stochastic declustering of earthquakes with the spatiotemporal renewal ETAS model." https://doi.org/10.1214/23-AOAS1756SUPP
UTSU, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine* **30** 521–605.
VELASCO HERRERA, V. M., ROSSELLO, E. A., ORGEIRA, M. J., ARIONI, L., SOON, W., VELASCO, G., LA ROSIQUE-DE, C. L., ZÚÑIGA, E. and VERA, C. (2022). Long-term forecasting of strong earthquakes in North America, South America, Japan, southern China and northern India with machine learning. *Front. Earth Sci.* **10**.
WAND, M. P. and JONES, M. C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* **9** 97–116. MR1280754
WHEATLEY, S., FILIMONOV, V. and SORNETTE, D. (2016). The Hawkes process with renewal immigration & its estimation with an EM algorithm. *Comput. Statist. Data Anal.* **94** 120–135. MR3412815 https://doi.org/10.1016/j.csda.2015.08.007
ZHUANG, J. and MATEU, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *J. Roy. Statist. Soc. Ser. A* **182** 919–942. MR3955503
ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459 https://doi.org/10.1198/016214502760046925
ZIPKIN, J. R., SCHOENBERG, F. P., CORONGES, K. and BERTOZZI, A. L. (2016). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European J. Appl. Math.* **27** 502–529. MR3491509 https://doi.org/10.1017/S0956792515000492

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3195–3215
https://doi.org/10.1214/23-AOAS1757

OPTIMAL SAMPLING DESIGNS FOR MULTIDIMENSIONAL STREAMING TIME SERIES WITH APPLICATION TO POWER GRID SENSOR DATA

BY RUI XIE[1,a], SHUYANG BAI[2,b] AND PING MA[2,c]

[1]*Department of Statistics and Data Science, University of Central Florida,* [a]*rui.xie@ucf.edu*
[2]*Department of Statistics, University of Georgia,* [b]*bsy9142@uga.edu,* [c]*pingma@uga.edu*

The Internet of Things (IoT) system generates massive high-speed temporally correlated streaming data and is often connected with online inference tasks under computational or energy constraints. Online analysis of these streaming time series data often faces a trade-off between statistical efficiency and computational cost. One important approach to balance this trade-off is sampling, where only a small portion of the sample is selected for the model fitting and update. Motivated by the demands of dynamic relationship analysis of IoT system, we study the data-dependent sample selection and online inference problem for a multidimensional streaming time series, aiming to provide low-cost real-time analysis of high-speed power grid electricity consumption data. Inspired by D-optimality criterion in design of experiments, we propose a class of online data reduction methods that achieve an optimal sampling criterion and improve the computational efficiency of the online analysis. We show that the optimal solution amounts to a strategy that is a mixture of Bernoulli sampling and leverage score sampling. The leverage score sampling involves auxiliary estimations that have a computational advantage over recursive least squares updates. Theoretical properties of the auxiliary estimations involved are also discussed. When applied to European power grid consumption data, the proposed leverage score based sampling methods outperform the benchmark sampling method in online estimation and prediction. The general applicability of the sampling-assisted online estimation method is assessed via simulation studies.

REFERENCES

AGARWAL, P. K., HAR-PELED, S. and VARADARAJAN, K. R. (2005). Geometric approximation via coresets. In *Combinatorial and Computational Geometry*. *Math. Sci. Res. Inst. Publ.* **52** 1–30. Cambridge Univ. Press, Cambridge. MR2178310 https://doi.org/10.4171/PRIMS/172

AKBAR, A., KHAN, A., CARREZ, F. and MOESSNER, K. (2017). Predictive analytics for complex IoT data streams. *IEEE Int. Things J.* **4** 1571–1582.

ANAGNOSTOPOULOS, C., HADJIEFTHYMIADES, S., KATSIKIS, A. and MAGLOGIANNIS, I. (2014). Autoregressive energy-efficient context forwarding in wireless sensor networks for pervasive healthcare systems. *Pers. Ubiquitous Comput.* **18** 101–114.

BALDUIN, S., VEITH, E. and LEHNHOFF, S. (2022). Sampling strategies for static powergrid models. arXiv preprint. Available at arXiv:2204.09053.

BERBERIDIS, D., KEKATOS, V. and GIANNAKIS, G. B. (2016). Online censoring for large-scale regressions with application to streaming big data. *IEEE Trans. Signal Process.* **64** 3854–3867. MR3515721 https://doi.org/10.1109/TSP.2016.2546225

BINGHAM, N. H., GOLDIE, C. M. and TEUGELS, J. L. (1989). *Regular Variation*. *Encyclopedia of Mathematics and Its Applications* **27**. Cambridge Univ. Press, Cambridge. MR1015093

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. and LJUNG, G. M. (2016). *Time Series Analysis*: *Forecasting and Control*, 5th ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3379415

CAI, D., SHI, D. and CHEN, J. (2013). Probabilistic load flow computation with polynomial normal transformation and Latin hypercube sampling. *IET Gener. Transm. Distrib.* **7** 474–482.

COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19** 15–18. MR0436478 https://doi.org/10.2307/1268249

Key words and phrases. Streaming data, sampling design, multidimensional time series, D-optimality.

DASGUPTA, A., DRINEAS, P., HARB, B., KUMAR, R. and MAHONEY, M. W. (2009). Sampling algorithms and coresets for l_p regression. *SIAM J. Comput.* **38** 2060–2078. MR2476287 https://doi.org/10.1137/070696507

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

ESHRAGH, A., ROOSTA, F., NAZARI, A. and MAHONEY, M. W. (2022). LSAR: Efficient leverage score sampling algorithm for the analysis of big time series data. *J. Mach. Learn. Res.* **23** Paper No. [22], 36. MR4420747

FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. *Monographs on Statistics and Applied Probability* **36**. CRC Press, London. MR1071174 https://doi.org/10.1007/978-1-4899-2937-2

FELDMAN, D., SCHMIDT, M. and SOHLER, C. (2012). Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms* 1434–1453. SIAM, Philadelphia, PA. MR3202989

GABEL, M., KEREN, D. and SCHUSTER, A. (2015). Monitoring least squares models of distributed streams. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 319–328. ACM, New York.

GIRAITIS, L., KOUL, H. L. and SURGAILIS, D. (2012). *Large Sample Inference for Long Memory Processes*. Imperial College Press, London. MR2977317 https://doi.org/10.1142/p591

GITTENS, A. and MAHONEY, M. (2013). Revisiting the Nyström method for improved large-scale machine learning. In *International Conference on Machine Learning*, 567–575. PMLR, Atlanta, GA, USA.

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

HILL, D. J. and MINSKER, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environ. Model. Softw.* **25** 1014–1022.

HOOI, B., SONG, H. A., PANDEY, A., JEREMINOV, M., PILEGGI, L. and FALOUTSOS, C. (2018). Streamcast: Fast and online mining of power grid time sequences. In *Proceedings of the* 2018 *SIAM International Conference on Data Mining* 531–539. SIAM, Philadelphia.

ISLAM, S. R., KWAK, D., KABIR, M. H., HOSSAIN, M. and KWAK, K.-S. (2015). The Internet of things for health care: A comprehensive survey. *IEEE Access* **3** 678–708.

JARADAT, M., JARRAH, M., BOUSSELHAM, A., JARARWEH, Y. and AL-AYYOUB, M. (2015). The Internet of energy: Smart sensor networks and big data management for smart grid. *Proc. Comput. Sci.* **56** 592–597.

JORDAN, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390. MR3102908 https://doi.org/10.3150/12-BEJSP17

JUMAR, R., MAASS, H., SCHÄFER, B., GORJÃO, L. R. and HAGENMEYER, V. (2020). Database of power grid frequency measurements. arXiv preprint. Available at arXiv:2006.01771.

KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1876169 https://doi.org/10.1007/978-1-4757-4015-8

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82** 35–45. MR3931993

KALMAN, R. E. and BUCY, R. S. (1961). New results in linear filtering and prediction theory. *J. Basic Eng.* **83** 95–108. MR0234760

LI, F., XIE, R., WANG, Z., GUO, L., YE, J., MA, P. and SONG, W. (2019). Online distributed IoT security monitoring with multidimensional streaming big data. *IEEE Int. Things J.* **7** 4387–4394.

LIBERTY, E. (2013). Simple and deterministic matrix sketching. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 581–588. ACM, New York.

LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. *Springer Series in Statistics*. Springer, New York. MR2401592

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

MA, P., CHEN, Y., ZHANG, X., XING, X., MA, J. and MAHONEY, M. W. (2022). Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *J. Mach. Learn. Res.* **23** 1–45.

MA, P., MAHONEY, M. W. and YU, B. (2015). A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.* **16** 861–911. MR3361306

MARANGONI, G. and TAVONI, M. (2021). Real-time feedback on electricity consumption: Evidence from a field experiment in Italy. *Energy Effic.* **14** 1–17.

MAT, I., KASSIM, M. R. M., HARUN, A. N. and YUSOFF, I. M. (2016). IoT in precision agriculture applications using wireless moisture sensor network. In 2016 *IEEE Conference on Open Systems* (*ICOS*) 24–29. IEEE, New York.

MENG, C., XIE, R., MANDAL, A., ZHANG, X., ZHONG, W. and MA, P. (2021). LowCon: A design-based subsampling approach in a misspecified linear model. *J. Comput. Graph. Statist.* **30** 694–708. MR4313470 https://doi.org/10.1080/10618600.2020.1844215

MICHALAREAS, G., SCHOFFELEN, J.-M., PATERSON, G. and GROSS, J. (2013). Investigating causality between interacting brain areas with multivariate autoregressive models of MEG sensor data. *Hum. Brain Mapp.* **34** 890–913.

NELLORE, K. and HANCKE, G. P. (2016). A survey on urban traffic management system using wireless sensor networks. *Sensors* **16** 157.

OPEN POWER SYSTEM DATA (2020). Data package time series. Version 2020-10-06. (Primary data from various sources. Available at https://doi.org/10.25832/time_series/2020-10-06.)

PAPALAMBROS, P. Y. and WILDE, D. J. (2000). *Principles of Optimal Design*: *Modeling and Computation*, 2nd ed. Cambridge Univ. Press, Cambridge. MR1775704 https://doi.org/10.1017/CBO9780511626418

PETRIS, G., PETRONE, S. and CAMPAGNOLI, P. (2009). *Dynamic Linear Models with R. Use R!* Springer, New York. MR2730074 https://doi.org/10.1007/b135794

PLACKETT, R. L. (1950). Some theorems in least squares. *Biometrika* **37** 149–157. MR0036980 https://doi.org/10.1093/biomet/37.1-2.149

PRONZATO, L. (2006). On the sequential construction of optimum bounded designs. *J. Statist. Plann. Inference* **136** 2783–2804. MR2279835 https://doi.org/10.1016/j.jspi.2004.10.020

PRONZATO, L. and WANG, H. (2021). Sequential online subsampling for thinning experimental designs. *J. Statist. Plann. Inference* **212** 169–193. MR4180110 https://doi.org/10.1016/j.jspi.2020.08.001

PUKELSHEIM, F. (2006). *Optimal Design of Experiments*. *Classics in Applied Mathematics* **50**. SIAM, Philadelphia, PA. Reprint of the 1993 original. MR2224698 https://doi.org/10.1137/1.9780898719109

QIU, H., XU, S., HAN, F., LIU, H. and CAFFO, B. (2015). Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *International Conference on Machine Learning* 1843–1851.

RÉMILLARD, B., PAPAGEORGIOU, N. and SOUSTRA, F. (2012). Copula-based semiparametric models for multivariate time series. *J. Multivariate Anal.* **110** 30–42. MR2927508 https://doi.org/10.1016/j.jmva.2012.03.001

SCHIMBINSCHI, F., MOREIRA-MATIAS, L., NGUYEN, V. X. and BAILEY, J. (2017). Topology-regularized universal vector autoregression for traffic forecasting in large urban areas. *Expert Syst. Appl.* **82** 301–316.

SEBER, G. A. and LEE, A. J. (2012). *Linear Regression Analysis*, *Vol.* 329. Wiley, New York.

SHEHABI, A., SMITH, S., SARTOR, D., BROWN, R., HERRLIN, M., KOOMEY, J., MASANET, E., HORNER, N., AZEVEDO, I. et al. (2016). United States data center energy usage report.

SHERMAN, J. and MORRISON, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Stat.* **21** 124–127. MR0035118 https://doi.org/10.1214/aoms/1177729893

SIDDIK, M. A. B., SHEHABI, A. and MARSTON, L. (2021). The environmental footprint of data centers in the United States. *Environ. Res. Lett.* **16** 064017.

TING, D. and BROCHU, E. (2018). Optimal subsampling with influence functions. In *Advances in Neural Information Processing Systems* 3654–3663.

WANG, H., YANG, M. and STUFKEN, J. (2019). Information-based optimal subdata selection for big data linear regression. *J. Amer. Statist. Assoc.* **114** 393–405. MR3941263 https://doi.org/10.1080/01621459.2017.1408468

WANG, H., ZHU, R. and MA, P. (2018). Optimal subsampling for large sample logistic regression. *J. Amer. Statist. Assoc.* **113** 829–844. MR3832230 https://doi.org/10.1080/01621459.2017.1292914

WANG, L., ELMSTEDT, J., WONG, W. K. and XU, H. (2021). Orthogonal subsampling for big data linear regression. *Ann. Appl. Stat.* **15** 1273–1290. MR4316648 https://doi.org/10.1214/21-aoas1462

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

WOODRUFF, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.* **10** 1–157. MR3285427 https://doi.org/10.1561/0400000060

XIE, R., BAI, S. and MA, P. (2023). Supplement to "Optimal sampling designs for multidimensional streaming time series with application to power grid sensor data." https://doi.org/10.1214/23-AOAS1757SUPP

XIE, R., WANG, Z., BAI, S., MA, P. and ZHONG, W. (2019). Online decentralized leverage score sampling for streaming multidimensional time series. In *The 22nd International Conference on Artificial Intelligence and Statistics* 2301–2311.

XU, X., CHEN, Y., GOUDE, Y. and YAO, Q. (2021). Day-ahead probabilistic forecasting for French half-hourly electricity loads and quantiles for curve-to-curve regression. *Appl. Energy* **301** 117465.

YOKOYAMA, R. (1980). Moment bounds for stationary mixing sequences. *Z. Wahrsch. Verw. Gebiete* **52** 45–57. MR0568258 https://doi.org/10.1007/BF00534186

YU, J., WANG, H., AI, M. and ZHANG, H. (2022). Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *J. Amer. Statist. Assoc.* **117** 265–276. MR4399084 https://doi.org/10.1080/01621459.2020.1773832

ZHANG, K., LIU, C., ZHANG, J., XIONG, H., XING, E. and YE, J. (2017). Randomization or condensation?: Linear-cost matrix sketching via cascaded compression sampling. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 615–623. ACM, New York.

ZHANG, T. and WU, W. B. (2012). Inference of time-varying regression models. *Ann. Statist.* **40** 1376–1402. MR3015029 https://doi.org/10.1214/12-AOS1010

ZHOU, B. and SAAD, W. (2019). Joint status sampling and updating for minimizing age of information in the Internet of things. *IEEE Trans. Commun.* **67** 7468–7482.

ZHOU, Z. and WU, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 513–531. MR2758526 https://doi.org/10.1111/j.1467-9868.2010.00743.x

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3216–3240
https://doi.org/10.1214/23-AOAS1758

A RIEMANN MANIFOLD MODEL FRAMEWORK FOR LONGITUDINAL CHANGES IN PHYSICAL ACTIVITY PATTERNS

BY JINGJING ZOU[1,6,a], TUO LIN[1,b], CHONGZHI DI[2,g], JOHN BELLETTIERE[1,c], MARTA M. JANKOWSKA[3,h], SHERI J. HARTMAN[1,6,d], DOROTHY D. SEARS[4,5,6,i], ANDREA Z. LACROIX[1,e], CHERYL L. ROCK[5,j] AND LOKI NATARAJAN[1,6,f]

[1]*Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego,* [a]*j2zou@health.ucsd.edu,* [b]*tulin@health.ucsd.edu,* [c]*jbellettiere@health.ucsd.edu,* [d]*sjhartman@health.ucsd.edu,* [e]*alacroix@health.ucsd.edu,* [f]*lnatarajan@health.ucsd.edu*

[2]*Division of Public Health Sciences, Fred Hutchinson Cancer Center,* [g]*cdi@fredhutch.org*

[3]*Department of Population Sciences, Beckman Research Institute, City of Hope,* [h]*mjankowska@coh.org*

[4]*College of Health Solutions, Arizona State University,* [i]*dsears@health.ucsd.edu*

[5]*Department of Family Medicine, University of California, San Diego,* [j]*clrock@health.ucsd.edu*

[6]*UC San Diego Moores Cancer Center*

Physical activity (PA) is significantly associated with many health outcomes. The wide usage of wearable accelerometer-based activity trackers in recent years has provided a unique opportunity for in-depth research on PA and its relations with health outcomes and interventions. Past analysis of activity tracker data relies heavily on aggregating minute-level PA records into day-level summary statistics in which important information of PA temporal/diurnal patterns is lost. In this paper we propose a novel functional data analysis approach based on Riemann manifolds for modeling PA and its longitudinal changes. We model smoothed minute-level PA of a day as one-dimensional Riemann manifolds and longitudinal changes in PA in different visits as deformations between manifolds. The variability in changes of PA among a cohort of subjects is characterized via variability in the deformation. Functional principal component analysis is further adopted to model the deformations, and PC scores are used as a proxy in modeling the relation between changes in PA and health outcomes and/or interventions. We conduct comprehensive analyses on data from two clinical trials: Reach for Health (RfH) and Metabolism, Exercise and Nutrition at UCSD (MENU), focusing on the effect of interventions on longitudinal changes in PA patterns and how different modes of changes in PA influence weight loss, respectively. The proposed approach reveals unique modes of changes, including overall enhanced PA, boosted morning PA, and shifts of active hours specific to each study cohort. The results bring new insights into the study of longitudinal changes in PA and health and have the potential to facilitate designing of effective health interventions and guidelines.

REFERENCES

ADAMO, K. B., PRINCE, S. A., TRICCO, A. C., CONNOR GORBER, S. and TREMBLAY, M. (2009). A comparison of indirect versus direct measures for assessing physical activity in the pediatric population: A systematic review. *Int. J. Pediatr. Obes.* **4** 2–27.

AINSWORTH, B., CAHALIN, L., BUMAN, M. and ROSS, R. (2014). The current state of physical activity assessment tools. *Prog. Cardiovasc. Dis.* **57** 387–395.

AMAGASA, S., MACHIDA, M., FUKUSHIMA, N., KIKUCHI, H., TAKAMIYA, T., ODAGIRI, Y. and INOUE, S. (2018). Is objectively measured light-intensity physical activity associated with health outcomes after adjustment for moderate-to-vigorous physical activity in adults? A systematic review. *Int. J. Behav. Nutr. Phys. Act.* **15** 1–13.

Key words and phrases. Activity trackers, accelerometer, functional data analysis, Riemann manifold, longitudinal analysis, functional principal component analysis.

ANIRUDH, R., TURAGA, P., SU, J. and SRIVASTAVA, A. (2016). Elastic functional coding of Riemannian trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 922–936.

BASSETT, D. R. (2012). Device-based monitoring in physical activity and public health research. *Physiol. Meas.* **33** 1769–1783.

BEG, M. F., MILLER, M. I., TROUVE, A. and YOUNES, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vis.* **61** 139–157.

BELLETTIERE, J., HEALY, G. N., LAMONTE, M. J., KERR, J., EVENSON, K. R., RILLAMAS-SUN, E., DI, C., BUCHNER, D. M., HOVELL, M. F. et al. (2019a). Sedentary behavior and prevalent diabetes in 6,166 older women: The objective physical activity and cardiovascular health study. *J. Gerontol., Ser. A, Biol. Sci. Med. Sci.* **74** 387–395.

BELLETTIERE, J., LAMONTE, M. J., EVENSON, K. R., RILLAMAS-SUN, E., KERR, J., LEE, I. M., DI, C., ROSENBERG, D. E., STEFANICK, M. L. et al. (2019b). Sedentary behavior and cardiovascular disease in older women: The opach study. *Circulation* **139** 1036–1046.

BELLETTIERE, J., LAMONTE, M. J., UNKART, J., LILES, S., LADDU-PATEL, D., MANSON, J. E., BANACK, H., SEGUIN-FOWLER, R., CHAVEZ, P. et al. (2020). Short physical performance battery and incident cardiovascular events among older women. *J. Amer. Heart Assoc.* **9** e016845. https://doi.org/10.1161/JAHA.120.016845

CHARLIER, B., CHARON, N. and TROUVÉ, A. (2017). The Fshape framework for the variability analysis of functional shapes. *Found. Comput. Math.* **17** 287–357. MR3627451 https://doi.org/10.1007/s10208-015-9288-2

CHASTIN, S. F. M., DE CRAEMER, M., DE COCKER, K., POWELL, L., VAN CAUWENBERG, J., DALL, P., HAMER, M. and STAMATAKIS, E. (2019). How does light-intensity physical activity associate with adult cardiometabolic health and mortality? Systematic review with meta-analysis of experimental and observational studies. *Br. J. Sports Med.* **53** 370–376.

CHOI, H., WANG, Q., TOLEDO, M., TURAGA, P., BUMAN, M. and SRIVASTAVA, A. (2018). Temporal alignment improves feature quality: An experiment on activity recognition with accelerometer data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* 349–357.

COLLEY, R. C., GARRIGUET, D., JANSSEN, I., CRAIG, C. L., CLARKE, J. and TREMBLAY, M. S. (2011). Physical activity of Canadian adults: Accelerometer results from the 2007 to 2009 Canadian health measures survey. *Health Rep.* **22** 7–14.

CRAINICEANU, C. M., STAICU, A.-M. and DI, C.-Z. (2009). Generalized multilevel functional regression. *J. Amer. Statist. Assoc.* **104** 1550–1561. MR2750578 https://doi.org/10.1198/jasa.2009.tm08564

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

DYRSTAD, S. M., HANSEN, B. H., HOLME, I. M. and ANDERSSEN, S. A. (2014). Comparison of self-reported versus accelerometer-measured physical activity. *Med. Sci. Sports Exerc.* **46** 99–106.

DZIAK, J. and SHIYKO, M. (2021). funreg: Functional regression for irregularly timed data. R package version 1.2.2.

EKELUND, U., TARP, J., STEENE-JOHANNESSEN, J., HANSEN, B. H., JEFFERIS, B., FAGERLAND, M. W., WHINCUP, P., DIAZ, K. M., HOOKER, S. P. et al. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: Systematic review and harmonised meta-analysis. *BMJ* **366** l4570.

FÜZÉKI, E., ENGEROFF, T. and BANZER, W. (2017). Health benefits of light-intensity physical activity: A systematic review of accelerometer data of the national health and nutrition examination survey (NHANES). *Sports Med.* **47** 1769–1793.

GAJARDO, A., CARROLL, C., CHEN, Y., DAI, X., FAN, J., HADJIPANTELIS, P. Z., HAN, K., JI, H., MUELLER, H.-G. et al. (2021). Fdapace: Functional data analysis and empirical dynamics. R package version 0.5.7.

GLASS, N. L., BELLETTIERE, J., JAIN, P., LAMONTE, M. J., LACROIX, A. Z. and WOMEN'S HEALTH INITIATIVE (2021). Evaluation of light physical activity measured by accelerometry and mobility disability during a 6-year follow-up in older women. *JAMA Netw. Open* **4** e210005. https://doi.org/10.1001/jamanetworkopen.2021.0005

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

GOLDSMITH, J., LIU, X., JACOBSON, J. S. and RUNDLE, A. (2016). New insights into activity patterns in children, found using functional data analyses. *Med. Sci. Sports Exerc.* **48** 1723–1729.

GOLDSMITH, J., ZIPUNNIKOV, V. and SCHRACK, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **71** 344–353. MR3366239 https://doi.org/10.1111/biom.12278

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

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

HERNANDEZ, M., BOSSA, M. N. and OLMOS, S. (2009). Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. *Int. J. Comput. Vis.* **85** 291–306.

KHAN, W. A. A., JACKSON, M. L., KENNEDY, G. A. and CONDUIT, R. (2021). A field investigation of the relationship between rotating shifts, sleep, mental health and physical activity of Australian paramedics. *Sci. Rep.* **11** 1–11.

KURTEK, S. (2017). A geometric approach to pairwise Bayesian alignment of functional data using importance sampling. *Electron. J. Stat.* **11** 502–531. MR3619315 https://doi.org/10.1214/17-EJS1243

LACROIX, A. Z., BELLETTIERE, J., RILLAMAS-SUN, E., DI, C., EVENSON, K. R., LEWIS, C. E., BUCHNER, D. M., STEFANICK, M. L., LEE, I.-M. et al. (2019). Association of light physical activity measured by accelerometry and incidence of coronary heart disease and cardiovascular disease in older women. *JAMA Netw. Open* **2** e190419.

LAMONTE, M. J., BUCHNER, D. M., RILLAMAS-SUN, E., DI, C., EVENSON, K. R., BELLETTIERE, J., LEWIS, C. E., LEE, I.-M., TINKER, L. F. et al. (2018). Accelerometer-measured physical activity and mortality in women aged 63 to 99. *J. Amer. Geriatr. Soc.* **66** 886–894.

LE, T., FLATT, S. W., NATARAJAN, L., PAKIZ, B., QUINTANA, E. L., HEATH, D. D., RANA, B. K. and ROCK, C. L. (2016). Effects of diet composition and insulin resistance status on plasma lipid levels in a weight loss intervention in women. *J. Amer. Heart Assoc.*

LI, H., KOZEY KEADLE, S., STAUDENMAYER, J., ASSAAD, H., HUANG, J. Z. and CARROLL, R. J. (2015). Methods to assess an exercise intervention trial based on 3-level functional data. *Biostatistics* **16** 754–771. MR3449841 https://doi.org/10.1093/biostatistics/kxv015

LI, H., STAUDENMAYER, J. and CARROLL, R. J. (2014). Hierarchical functional data with mixed continuous and binary measurements. *Biometrics* **70** 802–811. MR3295741 https://doi.org/10.1111/biom.12211

LOPRINZI, P. D. (2016). Light-intensity physical activity and all-cause mortality. *Am. J. Health Promot.* **31** 340–342.

MARRON, J. S., RAMSAY, J. O., SANGALLI, L. M. and SRIVASTAVA, A. (2015). Functional data analysis of amplitude and phase variation. *Statist. Sci.* **30** 468–484. MR3432837 https://doi.org/10.1214/15-STS524

MATTHEWS, C. E., OCKENE, I. S., FREEDSON, P. S., ROSAL, M. C., MERRIAM, P. A. and HEBERT, J. R. (2002). Moderate to vigorous physical activity and risk of upper-respiratory tract infection. *Med. Sci. Sports Exerc.* **34** 1242–1248.

MIGUELES, J. H., CADENAS-SANCHEZ, C., EKELUND, U., NYSTRÖM, C. D., MORA-GONZALEZ, J., LÖF, M., LABAYEN, I., RUIZ, J. R. and ORTEGA, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Med.* **47** 1821–1845.

MONTARULI, A., GALASSO, L., CAUMO, A., CÈ, E., PESENTI, C., ROVEDA, E. and ESPOSITO, F. (2017). The circadian typology: The role of physical activity and melatonin. *Sport Sci. Health* **13** 469–476.

NADER, P. R., BRADLEY, R. H., HOUTS, R. M., MCRITCHIE, S. L. and O'BRIEN, M. (2008). Moderate-to-vigorous physical activity from ages 9 to 15 years. *JAMA* **300** 295–305.

PARADA, H., MCDONALD, E., BELLETTIERE, J., EVENSON, K. R., LAMONTE, M. J. and LACROIX, A. Z. (2020). Associations of accelerometer-measured physical activity and physical activity-related cancer incidence in older women: Results from the WHI OPACH study. *Br. J. Cancer* **122** 1409–1416.

PATTERSON, R. E., MARINAC, C. R., NATARAJAN, L., HARTMAN, S. J., CADMUS-BERTRAM, L., FLATT, S. W., LI, H., PARKER, B., ORATOWSKI-COLEMAN, J. et al. (2016). Recruitment strategies, design, and participant characteristics in a trial of weight-loss and metformin in breast cancer survivors. *Contemp. Clin. Trials* **47** 64–71.

PATTERSON, R. E., MARINAC, C. R., SEARS, D. D., KERR, J., HARTMAN, S. J., CADMUS-BERTRAM, L., VILLASEÑOR, A., FLATT, S. W., GODBOLE, S. et al. (2018). The effects of metformin and weight loss on biomarkers associated with breast cancer outcomes. *J. Natl. Cancer Inst.* **110** 1239–1247.

PRINCE, S. A., ADAMO, K. B., HAMEL, M., HARDT, J., CONNOR GORBER, S. and TREMBLAY, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. *Int. J. Behav. Nutr. Phys. Act.* **5** 56–24.

RAMAKRISHNAN, R., DOHERTY, A., SMITH-BYRNE, K., RAHIMI, K., BENNETT, D., WOODWARD, M., WALMSLEY, R. and DWYER, T. (2021). Accelerometer measured physical activity and the incidence of cardiovascular disease: Evidence from the uk biobank cohort study. *PLoS Med.* **18**.

REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. MR2411660 https://doi.org/10.1198/016214507000000527

REUTER, C., BELLETTIERE, J., LILES, S., DI, C., SEARS, D. D., LAMONTE, M. J., STEFANICK, M. L., LACROIX, A. Z. and NATARAJAN, L. (2020). Diurnal patterns of sedentary behavior and changes in physical function over time among older women: A prospective cohort study. 1–11.

ROCK, C. L., FLATT, S. W., PAKIZ, B., QUINTANA, E. L., HEATH, D. D., RANA, B. K. and NATARAJAN, L. (2016). Effects of diet composition on weight loss, metabolic factors and biomarkers in a 1-year weight loss intervention in obese women examined by baseline insulin resistance status. *Metab. Clin. Exper.* **65** 1605–1613.

SHOU, H., ZIPUNNIKOV, V., CRAINICEANU, C. M. and GREVEN, S. (2015). Structured functional principal component analysis. *Biometrics* **71** 247–257. MR3335369 https://doi.org/10.1111/biom.12236

SRIVASTAVA, A., WU, W., KURTEK, S., KLASSEN, E. and MARRON, J. S. (2011). Registration of functional data using fisher-rao metric. arXiv preprint arXiv:1103.3817.

STAMATAKIS, E., GALE, J., BAUMAN, A., EKELUND, U., HAMER, M. and DING, D. (2019). Sitting time, physical activity, and risk of mortality in adults. *J. Am. Coll. Cardiol.* **73** 2062–2072. https://doi.org/10.1016/j.jacc.2019.02.031

R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

VAILLANT, M., MILLER, M. I., YOUNES, L. and TROUVÉ, A. (2004). Statistics on diffeomorphisms via tangent space representations. *NeuroImage* **23** S161–S169.

WALKER, R. L., GREENWOOD-HICKMAN, M. A., BELLETTIERE, J., LACROIX, A. Z., WING, D., HIGGINS, M., RICHMIRE, K. R., LARSON, E. B., CRANE, P. K. et al. (2021). Associations between physical function and device-based measures of physical activity and sedentary behavior patterns in older adults: Moving beyond moderate-to-vigorous intensity physical activity. *BMC Geriatr.* **21**.

WROBEL, J., BAUER, A., MCDONNELL, E. and GOLDSMITH, J. (2022). registr: Curve registration for exponential family functional data. R package version 2.1.0.

WROBEL, J., ZIPUNNIKOV, V., SCHRACK, J. and GOLDSMITH, J. (2019). Registration for exponential family functional data. *Biometrics* **75** 48–57. MR3953706 https://doi.org/10.1111/biom.12963

XIAO, L., HUANG, L., SCHRACK, J. A., FERRUCCI, L., ZIPUNNIKOV, V. and CRAINICEANU, C. M. (2015). Quantifying the lifetime circadian rhythm of physical activity: A covariate-dependent functional approach. *Biostatistics* **16** 352–367. MR3365433 https://doi.org/10.1093/biostatistics/kxu045

XU, S. Y., NELSON, S., KERR, J., GODBOLE, S., JOHNSON, E., PATTERSON, R. E., ROCK, C. L., SEARS, D. D., ABRAMSON, I. et al. (2019). Modeling temporal variation in physical activity using functional principal components analysis. *Stat. Biosci.* **11** 403–421.

ZOU, J., LIN, T., DI, C., BELLETTIERE, J., JANKOWSKA, M. M., HARTMAN, S. J., SEARS, D. D., LACROIX, A. Z., ROCK, C. L. and NATARAJAN, L. (2023). Supplement to "A Riemann manifold model framework for longitudinal changes in physical activity patterns." https://doi.org/10.1214/23-AOAS1758SUPP

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3241–3256
https://doi.org/10.1214/23-AOAS1759

A PENALIZED COMPLEXITY PRIOR FOR DEEP BAYESIAN TRANSFER LEARNING WITH APPLICATION TO MATERIALS INFORMATICS

BY MOHAMED A. ABBA[a], JONATHAN P. WILLIAMS[b] AND BRIAN J. REICH[c]

Department of Statistics, North Carolina State University, [a]*mabba@ncsu.edu,* [b]*jwilli27@ncsu.edu,* [c]*bjreich@ncsu.edu*

A key task in the emerging field of materials informatics is to use machine learning to predict a material's properties and functions. A fast and accurate predictive model allows researchers to more efficiently identify or construct a material with desirable properties. As in many fields, deep learning is one of the state-of-the art approaches, but fully training a deep learning model is not always feasible in materials informatics due to limitations on data availability, computational resources, and time. Accordingly, there is a critical need in the application of deep learning to materials informatics problems to develop efficient *transfer learning* algorithms. The Bayesian framework is natural for transfer learning because the model trained from the source data can be encoded in the prior distribution for the target task of interest. However, the Bayesian perspective on transfer learning is relatively unaccounted for in the literature and is complicated for deep learning because the parameter space is large and the interpretations of individual parameters are unclear. Therefore, rather than subjective prior distributions for individual parameters, we propose a new Bayesian transfer learning approach based on the penalized complexity prior on the Kullback–Leibler divergence between the predictive models of the source and target tasks. We show via simulations that the proposed method outperforms other transfer learning methods across a variety of settings. The proposed method is applied to predict the properties of a molecular crystal, based on its structural properties, and we show improved precision for estimating the band gap of a material compared to state-of-the-art methods currently used in materials science.

REFERENCES

ABBA, M. A., WILLIAMS, J. P. and REICH, B. J. (2023). Supplement to "A penalized complexity prior for deep Bayesian transfer learning with application to materials informatics." https://doi.org/10.1214/23-AOAS1759SUPPA, https://doi.org/10.1214/23-AOAS1759SUPPB

AGARAP, A. F. (2018). Deep learning using rectified linear units (RELU). Available at arXiv:1803.08375.

AHISHAKIYE, E., VAN GIJZEN, M. B., TUMWIINE, J., WARIO, R. and OBUNGOLOCH, J. (2021). A survey on deep learning in medical image reconstruction. *Intelligent Medicine*.

ALMALKI, S. J. and NADARAJAH, S. (2014). Modifications of the Weibull distribution: A review. *Reliab. Eng. Syst. Saf.* **124** 32–55.

BATRA, R. (2021). Accurate machine learning in materials science facilitated by using diverse data sources. *Nature* **589** 524–525. https://doi.org/10.1038/d41586-020-03259-4

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

BLUNDELL, C., CORNEBISE, J., KAVUKCUOGLU, K. and WIERSTRA, D. (2015). Weight Uncertainty in Neural Networks.

BUENO, A., BENITEZ, C., DE ANGELIS, S., MORENO, A. D. and IBÁÑEZ, J. M. (2019). Volcano-seismic transfer learning and uncertainty quantification with Bayesian neural networks. *IEEE Trans. Geosci. Remote Sens.* **58** 892–902.

CHANDRA, R. and KAPOOR, A. (2020). Bayesian neural multi-source transfer learning. *Neurocomputing* **378** 54–64.

CHEN, C., YE, W., ZUO, Y., ZHENG, C. and ONG, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31** 3564–3572.

Key words and phrases. Kullback–Leibler divergence, materials science, neural networks, variational Bayesian inference.

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

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

DARGAN, S., KUMAR, M., ROHIT AYYAGARI, M. and KUMAR, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. *Arch. Comput. Methods Eng.* **27** 1071–1092. MR4138449 https://doi.org/10.1007/s11831-019-09344-w

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 *IEEE Conference on Computer Vision and Pattern Recognition* 248–255. IEEE, New York.

DUBE, P., BHATTACHARJEE, B., PETIT-BOIS, E. and HILL, M. (2020). Improving transferability of deep neural networks. In *Domain Adaptation for Visual Understanding* 51–64. Springer, Berlin.

FORTUIN, V. (2022). Priors in Bayesian deep learning: A review. *Int. Stat. Rev.* **90** 563–591. MR4524825

GLOROT, X. and BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (*AISTATS*'10). *Society for Artificial Intelligence and Statistics*.

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

HIMANEN, L., GEURTS, A., FOSTER, A. S. and RINKE, P. (2019). Data-driven materials science: Status, challenges, and perspectives. *Adv. Sci.* **6** 1900808. https://doi.org/10.1002/advs.201900808

HUO, H. and RUPP, M. (2018). Unified representation of molecules and crystals for machine learning. Available at arXiv:1704.06439.

KARBALAYGHAREH, A., QIAN, X. and DOUGHERTY, E. R. (2018). Optimal Bayesian transfer learning. *IEEE Trans. Signal Process.* **66** 3724–3739. MR3841740 https://doi.org/10.1109/TSP.2018.2839583

KITTEL, C., MCEUEN, P. and WILEY, J. (2019). Introduction to solid state physics.

KOHN, W. and SHAM, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev.* (2) **140** A1133–A1138. MR0189732

LECUN, Y., BENGIO, Y. et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks* **3361** 1995.

MA, S., MA, Y., ZHANG, B., TIAN, Y. and JIN, Z. (2021). Forecasting system of computational time of DFT/TDDFT calculations under the multiverse ansatz via machine learning and cheminformatics. *ACS Omega* **6** 2001–2024.

MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *J. Mach. Learn. Res.* **17** Paper No. 81. MR3517104

MEDSKER, L. R. and JAIN, L. (2001). Recurrent neural networks. *Design and Applications* **5** 64–67.

O'HAGAN, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Saf.* **91** 1290–1300.

OLSTHOORN, B., GEILHUFE, R. M., BORYSOV, S. S. and BALATSKY, A. V. (2019). Band gap prediction for large organic crystal structures with machine learning. *Advanced Quantum Technologies* **2** 1900023.

PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

RAGHU, M., ZHANG, C., KLEINBERG, J. M. and BENGIO, S. (2019). Transfusion: Understanding Transfer Learning with Applications to Medical Imaging. *CoRR*. Available at arXiv:1902.07208.

SCHÜTT, K. T., KINDERMANS, P.-J., SAUCEDA, H. E., CHMIELA, S., TKATCHENKO, A. and MÜLLER, K.-R. (2017). Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. ArXiv preprint. Available at arXiv:1706.08566.

SENIOR, A. W., EVANS, R., JUMPER, J., KIRKPATRICK, J., SIFRE, L., GREEN, T., QIN, C., ŽÍDEK, A., NELSON, A. W. et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature* **577** 706–710.

SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

STUKE, A., KUNKEL, C., GOLZE, D., TODOROVIĆ, M., MARGRAF, J. T., REUTER, K., RINKE, P. and OBERHOFER, H. A. (2019). "OE62-dataset" of molecular orbital energies.

TAN, C., SUN, F., KONG, T., ZHANG, W., YANG, C. and LIU, C. (2018). A survey on deep transfer learning. In *International Conference on Artificial Neural Networks* 270–279. Springer, Berlin.

VAN DE WALLE, C. G. (2012). *Wide-Band-Gap Semiconductors*. Elsevier, Amsterdam.

VOULODIMOS, A., DOULAMIS, N., DOULAMIS, A. and PROTOPAPADAKIS, E. (2018). Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018** 7068349. https://doi.org/10.1155/2018/7068349

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 1–40.

WILSON, A. G. (2020). The case for Bayesian deep learning. ArXiv preprint. Available at arXiv:2001.10995.
WOHLERT, J., MUNK, A., SENGUPTA, S. and LAUMANN, F. (2018). Bayesian transfer learning for deep networks. *ViXra*.
YANG, H., JIAO, S. and SUN, P. (2020). Bayesian-convolutional neural network model transfer learning for image detection of concrete water-binder ratio. *IEEE Access* **8** 35350–35367.
YOUNG, T., HAZARIKA, D., PORIA, S. and CAMBRIA, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13** 55–75.
ZEILER, M. D. and FERGUS, R. (2013). Visualizing and Understanding Convolutional Networks. *CoRR*. Available at arXiv:1311.2901.
ZHANG, C., BUTEPAGE, J., KJELLSTROM, H. and MANDT, S. (2019). Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774
ZHOU, C., ZHANG, J., LIU, J., ZHANG, C., SHI, G. and HU, J. (2020). Bayesian transfer learning for object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **58** 7705–7719.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3257–3282
https://doi.org/10.1214/23-AOAS1760

A GENERAL FRAMEWORK FOR PENALIZED MIXED-EFFECTS MULTITASK LEARNING WITH APPLICATIONS ON DNA METHYLATION SURROGATE BIOMARKERS CREATION

BY ANDREA CAPPOZZO[1,a], FRANCESCA IEVA[1,b] AND GIOVANNI FIORITO[2,c]

[1]*MOX, Department of Mathematics, Politecnico di Milano,* [a]*andrea.cappozzo@polimi.it,* [b]*francesca.ieva@polimi.it*
[2]*Clinical Bioinformatics Unit, IRCCS Istituto Giannina Gaslini,* [c]*giovannifiorito@gaslini.org*

Recent evidence highlights the usefulness of DNA methylation (DNAm) biomarkers as surrogates for exposure to risk factors for noncommunicable diseases in epidemiological studies and randomized trials. DNAm variability has been demonstrated to be tightly related to lifestyle behavior and exposure to environmental risk factors, ultimately providing an unbiased proxy of an individual state of health. At present, the creation of DNAm surrogates relies on univariate penalized regression models, with elastic-net regularizer being the gold standard when accomplishing the task. Nonetheless, more advanced modeling procedures are required in the presence of multivariate outcomes with a structured dependence pattern among the study samples. In this work we propose a general framework for mixed-effects multitask learning in presence of high-dimensional predictors to develop a multivariate DNAm biomarker from a multicenter study. A penalized estimation scheme, based on an expectation-maximization algorithm, is devised in which any penalty criteria for fixed-effects models can be conveniently incorporated in the fitting process. We apply the proposed methodology to create novel DNAm surrogate biomarkers for multiple correlated risk factors for cardiovascular diseases and comorbidities. We show that the proposed approach, modeling multiple outcomes together, outperforms state-of-the-art alternatives both in predictive power and biomolecular interpretation of the results.

REFERENCES

ANASTASIADI, D., ESTEVE-CODINA, A. and PIFERRER, F. (2018). Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenet. Chromatin* **11** 37. https://doi.org/10.1186/s13072-018-0205-1

ATCHLEY, W. R. and HALL, B. K. (1991a). A model for development and evolution of complex morphological structures. *Biol. Rev. Camb. Philos. Soc.* **66** 101–157.

ATCHLEY, W. R. and HALL, B. K. (1991b). A model for development and evolution of complex morphological structures. *Biol. Rev.* **66** 101–157.

AZZALINI, A. and CAPITANIO, A. (2013). *The Skew-Normal and Related Families* **3**. Cambridge Univ. Press, Cambridge.

BATTRAM, T., YOUSEFI, P., CRAWFORD, G., PRINCE, C., BABAEI, M. S., SHARP, G., HATCHER, C., VEGA-SALAS, M. J., KHODABAKHSH, S. et al. (2022). The EWAS catalog: A database of epigenome-wide association studies. *Wellcome Open Res.* **7**.

CAMPAGNA, M. P., XAVIER, A., LECHNER-SCOTT, J., MALTBY, V., SCOTT, R. J., BUTZKUEVEN, H., JOKUBAITIS, V. G. and LEA, R. A. (2021). Epigenome-wide association studies: Current knowledge, strategies and recommendations. *Clin. Epigenet.* **13** 214. https://doi.org/10.1186/s13148-021-01200-8

CAPPOZZO, A., IEVA, F. and FIORITO, G. (2023). Supplement to "A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation." https://doi.org/10.1214/23-AOAS1760SUPPA, https://doi.org/10.1214/23-AOAS1760SUPPB

CAPPOZZO, A., MCCRORY, C., ROBINSON, O., FRENI STERRANTINO, A., SACERDOTE, C., KROGH, V., PANICO, S., TUMINO, R., IACOVIELLO, L. et al. (2022). A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clin. Epigenet.* **14** 121.

Key words and phrases. Mixed-effects models, multitask learning, EM algorithm, penalized estimation, multivariate regression, personalized medicine.

CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75.
CASTRO DE MOURA, M., DAVALOS, V., PLANAS-SERRA, L., ALVAREZ-ERRICO, D., ARRIBAS, C., RUIZ, M., AGUILERA-ALBESA, S., TROYA, J., VALENCIA-RAMOS, J. et al. (2021). Epigenome-wide association study of Covid-19 severity with respiratory failure. *eBioMedicine* **66** 103339.
CHENG, W., ZHANG, X., GUO, Z., SHI, Y. and WANG, W. (2014). Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics* **30** 139–148.
CHIPPERFIELD, J. O. and STEEL, D. G. (2012). Multivariate random effect models with complete and incomplete data. *J. Multivariate Anal.* **109** 146–155. MR2922860 https://doi.org/10.1016/j.jmva.2012.02.014
CHUNG, F. R. K. and GRAHAM, F. C. (1997). *Spectral Graph Theory* **92**. Am. Math. Soc., Providence.
COLICINO, E., JUST, A., KIOUMOURTZOGLOU, M.-A., VOKONAS, P., CARDENAS, A., SPARROW, D., WEISSKOPF, M., NIE, L. H., HU, H. et al. (2021). Blood DNA methylation biomarkers of cumulative lead exposure in adults. *J. Expo. Sci. Environ. Epidemiol.* **31** 108–116.
CONOLE, E. L. S., STEVENSON, A. J., GREEN, C., HARRIS, S. E., MANIEGA, S. M., VALDÉS-HERNÁNDEZ, M. D. C., HARRIS, M. A., BASTIN, M. E., WARDLAW, J. M. et al. (2020). An epigenetic proxy of chronic inflammation outperforms serum levels as a biomarker of brain ageing. *MedRxiv* 2020.10.08.20205245.
GENE ONTOLOGY CONSORTIUM (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** 258D–261.
DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265
DEBRUINE, L. (2021). faux: Simulation for Factorial Designs.
DEMIDENKO, E. (2013). *Mixed Models*: *Theory and Applications with R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235905
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
DIRMEIER, S., FUCHS, C., MUELLER, N. S. and THEIS, F. J. (2018). netReg: Network-regularized linear models for biological association studies. *Bioinformatics* **34** 896–898. https://doi.org/10.1093/bioinformatics/btx677
DONG, W., CHEN, H., WANG, L., CAO, X., BU, X., PENG, Y., DONG, A., YING, M., CHEN, X. et al. (2020). Exploring the shared genes of hypertension, diabetes and hyperlipidemia based on microarray. *Braz. J. Pharm. Sci.* **56** 1–12.
FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F. et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* **46** D649–D655.
FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x
FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038. MR2550099
FAZZARI, M. J. and GREALLY, J. M. (2010). Introduction to Epigenomics and Epigenome-Wide Analysis. In *Statistical Methods in Molecular Biology* 243–265. Humana Press, Totowa, NJ.
FERNÁNDEZ-SANLÉS, A., SAYOLS-BAIXERAS, S., SUBIRANA, I., SENTÍ, M., PÉREZ-FERNÁNDEZ, S., DE CASTRO MOURA, M., ESTELLER, M., MARRUGAT, J. and ELOSUA, R. (2021). DNA methylation biomarkers of myocardial infarction and cardiovascular disease. *Clin. Epigenet.* **13** 86. https://doi.org/10.1186/s13148-021-01078-6
FIORITO, G., PEDRON, S., OCHOA-ROSALES, C., MCCRORY, C., POLIDORO, S., ZHANG, Y., DUGUÉ, P.-A., RATLIFF, S., ZHAO, W. N. et al. (2022). The role of epigenetic clocks in explaining educational inequalities in mortality: A multicohort study and meta-analysis. *J. Gerontol., Ser. A* **77** 1750–1759.
FIORITO, G., VLAANDEREN, J., POLIDORO, S., GULLIVER, J., GALASSI, C., RANZI, A., KROGH, V., GRIONI, S., AGNOLI, C. et al. (2018). Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. *Environ. Mol. Mutagen.* **59** 234–246. https://doi.org/10.1002/em.22153
FROHLICH, H. and ZELL, A. (2005). Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In *Proceedings*. 2005 *IEEE International Joint Conference on Neural Networks*, 2005. **3** 1431–1436. IEEE, Los Alamitos.
GAŁECKI, A. and BURZYKOWSKI, T. (2013). *Linear Mixed-Effects Models Using R. Springer Texts in Statistics*. Springer, New York. MR3024843 https://doi.org/10.1007/978-1-4614-3900-4
GUIDA, F., SANDANGER, T. M., CASTAGNÉ, R., CAMPANELLA, G., POLIDORO, S., PALLI, D., KROGH, V., TUMINO, R., SACERDOTE, C. et al. (2015). Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24** 2349–2359. https://doi.org/10.1093/hmg/ddu751

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

HIDALGO, B. A., MINNIEFIELD, B., PATKI, A., TANNER, R., BAGHERI, M., TIWARI, H. K., ARNETT, D. K. and IRVIN, M. R. (2021). A 6-CpG validated methylation risk score model for metabolic syndrome: The HyperGEN and GOLDN studies. *PLoS ONE* **16** e0259836. https://doi.org/10.1371/journal.pone.0259836

HILLARY, R. F. and MARIONI, R. E. (2020). MethylDetectR: A software for methylation-based health profiling. *Wellcome Open Res*. **5** 283. https://doi.org/10.12688/wellcomeopenres.16458.2

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

JORDAN, M. I. (2013). On statistics, computation and scalability. *Bernoulli* **19** 1378–1390. MR3102908 https://doi.org/10.3150/12-BEJSP17

KIM, S., PAN, W. and SHEN, X. (2013). Network-based penalized regression with application to genomic data. *Biometrics* **69** 582–593. MR3106586 https://doi.org/10.1111/biom.12035

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to EQTL mapping. *Ann. Appl. Stat*. **6** 1095–1117. MR3012522 https://doi.org/10.1214/12-AOAS549

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform*. **9** 559.

LARIA, J. C., CARMEN AGUILERA-MORILLO, M. and LILLO, R. E. (2019). An iterative sparse-group lasso. *J. Comput. Graph. Statist*. **28** 722–731. MR4007753 https://doi.org/10.1080/10618600.2019.1573687

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat*. **4** 1498–1516. MR2758338 https://doi.org/10.1214/10-AOAS332

LI, Y., NAN, B. and ZHU, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* **71** 354–363. MR3366240 https://doi.org/10.1111/biom.12292

LOZA, M. J., MCCALL, C. E., LI, L., ISAACS, W. B., XU, J. and CHANG, B.-L. (2007). Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS ONE* **2** e1035.

LU, A. T., QUACH, A., WILSON, J. G., REINER, A. P., AVIV, A., RAJ, K., HOU, L., BACCARELLI, A. A., LI, Y. et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11** 303–327.

MARABITA, F., ALMGREN, M., LINDHOLM, M. E., RUHRMANN, S., FAGERSTRÖM-BILLAI, F., JAGODIC, M., SUNDBERG, C. J., EKSTRÖM, T. J., TESCHENDORFF, A. E. et al. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the illumina HumanMethylation450 BeadChip platform. *Epigenetics* **8** 333–346. https://doi.org/10.4161/epi.24008

MCCULLOCH, C. E. and NEUHAUS, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statist. Sci*. **26** 388–402. MR2917962 https://doi.org/10.1214/11-STS361

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2392878 https://doi.org/10.1002/9780470191613

MENG, X.-L. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503 https://doi.org/10.1093/biomet/80.2.267

NGUYEN, T. M., LE, H. L., HWANG, K.-B., HONG, Y.-C. and KIM, J. H. (2022). Predicting high blood pressure using DNA methylome-based machine learning models. *Biomedicines* **10** 1406.

OBOZINSKI, G., TASKAR, B. and JORDAN, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput*. **20** 231–252. MR2610775 https://doi.org/10.1007/s11222-008-9111-x

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2009). High-dimensional support union recovery in multivariate regression. In *Advances in Neural Information Processing Systems* 21—*Proceedings of the* 2008 *Conference* 1217–1224.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011b). Support union recovery in high-dimensional multivariate regression. *Ann. Statist*. **39** 1–47. MR2797839 https://doi.org/10.1214/09-AOS776

ODINTSOVA, V. V., REBATTU, V., HAGENBEEK, F. A., POOL, R., BECK, J. J., EHLI, E. A., VAN BEIJSTERVELDT, C. E. M., LIGTHART, L., WILLEMSEN, G. et al. (2021). Predicting complex traits and exposures from polygenic scores and blood and buccal DNA methylation profiles. *Front. Psychiatr.* **12** 1–17.

PANICO, S., DELLO IACOVO, R., CELENTANO, E., GALASSO, R., MUTI, P., SALVATORE, M. and MANCINI, M. (1992). Progetto ATENA, a study on the etiology of major chronic diseases in women: Design, rationale and objectives. *Eur. J. Epidemiol*. **8** 601–608.

PHIPSON, B., MAKSIMOVIC, J. and OSHLACK, A. (2016). missMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32** 286–288.

PINHEIRO, J. and BATES, D. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin.

RAULUSEVICIUTE, I., DRABLØS, F. and RYE, M. B. (2020). DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Med. Genom*. **13** 6.

REIMAND, J., ISSERLIN, R., VOISIN, V., KUCERA, M., TANNUS-LOPES, C., ROSTAMIANFAR, A., WADI, L., MEYER, M., WONG, J. et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14** 482–517.

REINSEL, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.* **79** 406–414. MR0755095

RIBOLI, E., HUNT, K., SLIMANI, N., FERRARI, P., NORAT, T., FAHEY, M., CHARRONDIÈRE, U., HÉMON, B., CASAGRANDE, C. et al. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): Study populations and data collection. *Public Health Nutr.* **5** 1113–1124.

RICHARD, M. A., HUAN, T., LIGTHART, S., GONDALIA, R., JHUN, M. A., BRODY, J. A., IRVIN, M. R., MARIONI, R., SHEN, J. et al. (2017). DNA methylation analysis identifies loci for blood pressure regulation. *Am. J. Hum. Genet.* **101** 888–902.

RODOSTHENOUS, T., SHAHREZAEI, V. and EVANGELOU, M. (2020). Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: A comparison study. *Bioinformatics* **36** 4616–4625. https://doi.org/10.1093/bioinformatics/btaa530

ROHART, F., SAN CRISTOBAL, M. and LAURENT, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Comput. Statist. Data Anal.* **80** 209–222. MR3240488 https://doi.org/10.1016/j.csda.2014.06.022

SCHAFER, J. L. and YUCEL, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.* **11** 437–457. MR1938143 https://doi.org/10.1198/106186002760180608

SCHELLDORFER, J., BÜHLMANN, P. and VAN DE GEER, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.* **38** 197–214. MR2829596 https://doi.org/10.1111/j.1467-9469.2011.00740.x

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

SENONER, T. and DICHTL, W. (2019). Oxidative stress in cardiovascular diseases: Still a therapeutic target? *Nutrients* **11**.

SHAH, A., LAIRD, N. and SCHOENFELD, D. (1997). A random-effects model for multiple characteristics with possibly missing data. *J. Amer. Statist. Assoc.* **92** 775–779. MR1467867 https://doi.org/10.2307/2965726

SIGRIST, F. (2022). Latent Gaussian model boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* 1–1.

SILL, M., HIELSCHER, T., BECKER, N. and ZUCKNICK, M. (2014). c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. *J. Stat. Softw.* **62**.

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

SINGAL, R. and GINDER, G. D. (1999). DNA methylation. *Blood* **93** 4059–4070.

STEVENSON, A. J., MCCARTNEY, D. L., HILLARY, R. F., CAMPBELL, A., MORRIS, S. W., BERMINGHAM, M. L., WALKER, R. M., EVANS, K. L., BOUTIN, T. S. et al. (2020). Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin. Epigenet.* **12** 113.

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

TAY, J. K., NARASIMHAN, B. and HASTIE, T. (2021). Elastic net regularization paths for all generalized linear models.

R CORE TEAM (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

TYLER, A. L., CRAWFORD, D. C. and PENDERGRASS, S. A. (2013). Detecting and characterizing pleiotropy: New methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes. In *Biocomputing* 2014 183–187. World Scientific, Singapore.

VAN EIJK, K. R., DE JONG, S., BOKS, M. P. M., LANGEVELD, T., COLAS, F., VELDINK, J. H., DE KOVEL, C. G. F., JANSON, E., STRENGMAN, E. et al. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13** 636.

VINGA, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. *Brief. Bioinform.* **22** 77–87. https://doi.org/10.1093/bib/bbaa122

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WITTEN, D. M. and TIBSHIRANI, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* **8** 28. MR2533636 https://doi.org/10.2202/1544-6115.1470

WU, C.-Y., HU, H.-Y., CHOU, Y.-J., HUANG, N., CHOU, Y.-C. and LI, C.-P. (2015). High blood pressure and all-cause and cardiovascular disease mortalities in community-dwelling older adults. *Medicine* **94** e2160.

YI, Y., FANG, Y., WU, K., LIU, Y. and ZHANG, W. (2020). Comprehensive gene and pathway analysis of cervical cancer progression. *Oncol. Lett.* **19** 3316–3332.

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

YUAN, T., EDELMANN, D., FAN, Z., ALWERS, E., KATHER, J. N., BRENNER, H. and HOFFMEISTER, M. (2022). Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: A systematic review of epigenome-wide studies. *MedRxiv*.

ZHANG, Y., ELGIZOULI, M., SCHÖTTKER, B., HOLLECZEK, B., NIETERS, A. and BRENNER, H. (2016). Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin. Epigenet.* **8** 1–12.

ZHAO, Z., BANTERLE, M., BOTTOLO, L., RICHARDSON, S., LEWIN, A. and ZUCKNICK, M. (2021a). BayesSUR: An R package for high-dimensional multivariate Bayesian variable and covariance selection in linear regression. *J. Stat. Softw.* **100**.

ZHAO, Z., BANTERLE, M., LEWIN, A. and ZUCKNICK, M. (2021b). Structured Bayesian variable selection for multiple related response variables and high-dimensional predictors. ArXiv Preprint. Available at arXiv:2101.05899, 1–33.

ZHAO, Z., WANG, S., ZUCKNICK, M. and AITTOKALLIO, T. (2022). Tissue-specific identification of multi-omics features for pan-cancer drug response prediction. *iScience* **25** 104767.

ZHAO, Z. and ZUCKNICK, M. (2020). Structured penalized regression for drug sensitivity prediction. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 525–545. MR4098960

ZHONG, J., AGHA, G. and BACCARELLI, A. A. (2016). The role of DNA methylation in cardiovascular risk and disease: Methodological aspects, study design, and data analysis for epidemiological studies. *Circ. Res.* **118** 119–131. https://doi.org/10.1161/CIRCRESAHA.115.305206

ZHONG, W., WANG, J. and CHEN, X. (2021). Censored mean variance sure independence screening for ultrahigh dimensional survival data. *Comput. Statist. Data Anal.* **159** 107206. MR4233350 https://doi.org/10.1016/j.csda.2021.107206

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 768–768.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3283–3299
https://doi.org/10.1214/23-AOAS1762

A DYNAMIC ADDITIVE AND MULTIPLICATIVE EFFECTS NETWORK MODEL WITH APPLICATION TO THE UNITED NATIONS VOTING BEHAVIORS

BY BOMIN KIM[1,a], XIAOYUE NIU[2,b], DAVID HUNTER[2,c] AND XUN CAO[3,d]

[1]*Economic and Housing Research, Freddie Mac,* [a]*bomin8319@gmail.com*
[2]*Department of Statistics, Pennsylvania State University,* [b]*xiaoyue@psu.edu,* [c]*dhunter@stat.psu.edu*
[3]*Department of Political Science, Pennsylvania State University,* [d]*xuc11@psu.edu*

Motivated by a study of United Nations voting behaviors, we introduce a regression model for a series of networks that are correlated over time. Our model is a dynamic extension of the additive and multiplicative effects network model (AMEN) of Hoff (*Statist. Sci.* **36** (2021) 34–50). In addition to incorporating a temporal structure, the model accommodates two types of missing data and thus allows the size of the network to vary over time. We demonstrate via simulations the necessity of various components of the model. We apply the model to the United Nations General Assembly voting data from 1983 to 2014 (In *Routledge Handbook of International Organization* (2013) Routledge) to answer interesting research questions regarding international voting behaviors. In addition to finding important factors that could explain the voting behaviors, the model-estimated additive effects, multiplicative effects, and their movements reveal meaningful foreign policy positions and alliances of various countries.

REFERENCES

BAILEY, M. A., STREZHNEV, A. and VOETEN, E. (2017). Estimating dynamic state preferences from United Nations voting data. *J. Confl. Resolut.* **61** 430–456.

DURANTE, D. and DUNSON, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101** 883–898.

FRIEL, N., RASTELLI, R., WYSE, J. and RAFTERY, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proc. Natl. Acad. Sci. USA* **113** 6629–6634.

GARTZKE, E. (1998). Kant we all just get along? Opportunity, willingness, and the origins of the democratic peace. *Amer. J. Polit. Sci.* **42** 1–27.

GARTZKE, E. (2000). Preferences and the democratic peace. *International Studies Quarterly* **44** 191–212.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533.

GIBLER, D. M. (2008). *International Military Alliances*, 1648–2008. CQ Press.

GOODREAU, S. M., KITTS, J. A. and MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46** 103–125.

HANNEKE, S., FU, W. and XING, E. P. (2010). Discrete temporal models of social networks. *Electron. J. Stat.* **4** 585–605.

HOFF, P. (2021). Additive and multiplicative effects network models. *Statist. Sci.* **36** 34–50.

HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295.

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.

HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258.

KIM, B., LEE, K. H., XUE, L. and NIU, X. (2018). A review of dynamic network models with latent variables. *Stat. Surv.* **12** 105–135.

KIM, B., NIU, X., HUNTER, D. and CAO, X. (2023). Supplement to "A dynamic additive and multiplicative effects network model with application to the United Nations voting behaviors." https://doi.org/10.1214/23-AOAS1762SUPPA, https://doi.org/10.1214/23-AOAS1762SUPPB

Key words and phrases. Latent space model, varying number of nodes, international policy.

KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46.

LEIBENSTEIN, H. (1966). Shaping the world economy: Suggestions for an international economic policy.

LUSHER, D., KOSKINEN, J. and ROBINS, G. (2013). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge.

MARSHALL, M. G., JAGGERS, K. and GURR, T. R. (2014). Polity IV annual time-series, 1800–2013. Center for International Development and Conflict Management at the University of Maryland College Park.

OLIVELLA, S., PRATT, T. and IMAI, K. (2022). Dynamic stochastic blockmodel regression for network data: Application to international militarized conflicts. *J. Amer. Statist. Assoc.* **117** 1068–1081.

RASMUSSEN, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning* 63–71. Springer, Berlin.

RODRIGUE, J.-P., COMTOIS, C. and SLACK, B. (2016). *The Geography of Transport Systems*. Routledge, London.

SEWELL, D. K. and CHEN, Y. (2015). Latent space models for dynamic networks. *J. Amer. Statist. Assoc.* **110** 1646–1657.

SIGNORINO, C. S. and RITTER, J. M. (1999). Tau-b or not tau-b: Measuring the similarity of foreign policy positions. *International Studies Quarterly* **43** 115–144.

SNIJDERS, T. A., VAN DE BUNT, G. G. and STEGLICH, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Soc. Netw.* **32** 44–60.

SNIJDERS, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociol. Method.* **31** 361–395.

VOETEN, E. (2013). Data and analyses of voting in the UN general assembly. In *Routledge Handbook of International Organization* (B. Reinalda, ed.). Routledge, London. Available at http://ssrn.com/abstract=2111149.

WARD, M. D. and HOFF, P. D. (2007). Persistent patterns of international commerce. *J. Peace Res.* **44** 157–175.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.). *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3300–3323
https://doi.org/10.1214/23-AOAS1763

SEQUENTIAL MONTE CARLO FOR SAMPLING BALANCED AND COMPACT REDISTRICTING PLANS

BY CORY MCCARTAN[a] AND KOSUKE IMAI[b]

Department of Statistics, Harvard University, [a]*cmccartan@g.harvard.edu,* [b]*imai@harvard.edu*

Random sampling of graph partitions under constraints has become a popular tool for evaluating legislative redistricting plans. Analysts detect partisan gerrymandering by comparing a proposed redistricting plan with an ensemble of sampled alternative plans. For successful application sampling methods must scale to maps with a moderate or large number of districts, incorporate realistic legal constraints, and accurately and efficiently sample from a selected target distribution. Unfortunately, most existing methods struggle in at least one of these areas. We present a new sequential Monte Carlo (SMC) algorithm that generates a sample of redistricting plans converging to a realistic target distribution. Because it draws many plans in parallel, the SMC algorithm can efficiently explore the relevant space of redistricting plans better than the existing Markov chain Monte Carlo (MCMC) algorithms that generate plans sequentially. Our algorithm can simultaneously incorporate several constraints commonly imposed in real-world redistricting problems, including equal population, compactness, and preservation of administrative boundaries. We validate the accuracy of the proposed algorithm by using a small map where all redistricting plans can be enumerated. We then apply the SMC algorithm to evaluate the partisan implications of several maps submitted by relevant parties in a recent high-profile redistricting case in the State of Pennsylvania. We find that the proposed algorithm converges faster and with fewer samples than a comparable MCMC algorithm. Open-source software is available for implementing the proposed methodology.

REFERENCES

AKITAYA, H. A., KORMAN, M., KORTEN, O., SOUVAINE, D. L. and TÓTH, C. D. (2022). Reconfiguration of connected graph partitions via recombination. *Theoret. Comput. Sci.* **923** 13–26. MR4436558 https://doi.org/10.1016/j.tcs.2022.04.049

AUTRY, E., CARTER, D., HERSCHLAG, G., HUNTER, Z. and MATTINGLY, J. (2020). Multi-scale merge-split Markov chain Monte Carlo for redistricting. Working paper.

BANGIA, S., GRAVES, C. V., HERSCHLAG, G., KANG, H. S., LUO, J., MATTINGLY, J. C. and RAVIER, R. (2017). Redistricting: Drawing the line. arXiv preprint arXiv:1704.03360.

BOZKAYA, B., ERKUT, E. and LAPORTE, G. (2003). A tabu search heuristic and adaptive memory procedure for political districting. *European J. Oper. Res.* **144** 12–26.

CANNON, S., DUCHIN, M., RANDALL, D. and RULE, P. (2022). Spanning tree methods for sampling graph partitions. arXiv preprint arXiv:2210.01401.

CARTER, D., HERSCHLAG, G., HUNTER, Z. and MATTINGLY, J. (2019). A merge-split proposal for reversible Monte Carlo Markov Chain sampling of redistricting plans. arXiv preprint arXiv:1911.01503.

CHATTERJEE, S. and DIACONIS, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.* **28** 1099–1135. MR3784496 https://doi.org/10.1214/17-AAP1326

CHEN, J. (2017). Expert report of Jowei Chen, Ph.D. Expert witness report in League of Women Voters v. Commonwealth.

CHEN, J. and RODDEN, J. (2013). Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Q. J. Polit. Sci.* **8** 239–269.

CHIKINA, M., FRIEZE, A. and PEGDEN, W. (2017). Assessing significance in a Markov chain without mixing. *Proc. Natl. Acad. Sci. USA* **114** 2860–2864. MR3628186 https://doi.org/10.1073/pnas.1617540114

CHIKINA, M., FRIEZE, A. and PEGDEN, W. (2019). Understanding our Markov chain significance test: A reply to Cho and Rubinstein-Salzedo. *Stat. Public Policy* **6** 50–53.

Key words and phrases. Gerrymandering, graph partition, sequential Monte Carlo, spanning trees.

CHO, W. K. T. and LIU, Y. Y. (2018). Sampling from complicated and unknown distributions: Monte Carlo and Markov Chain Monte Carlo methods for redistricting. *Phys. A* **506** 170–178. MR3810349 https://doi.org/10.1016/j.physa.2018.03.096

CHO, W. K. T. and RUBINSTEIN-SALZEDO, S. (2019). Understanding significance tests from a non-mixing Markov chain for partisan gerrymandering claims. *Stat. Public Policy* **6** 44–49.

CIRINCIONE, C., DARLING, T. A. and O'ROURKE, T. G. (2000). Assessing South Carolina's 1990s congressional districting. *Polit. Geogr.* **19** 189–211.

COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. MR2239987

DEFORD, D., DUCHIN, M. and SOLOMON, J. (2021). Recombination: A family of Markov chains for redistricting. *Harv. Data Sci. Rev.* https://doi.org/10.1162/99608f92.eb30390f

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333 https://doi.org/10.1111/j.1467-9868.2006.00553.x

DOUCET, A., DE FREITAS, N. and GORDON, N., eds. (2001). *Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science*. Springer, New York. MR1847783 https://doi.org/10.1007/978-1-4757-3437-9

DUBE, M. P. and CLARK, J. T. (2016). Beyond the circle: Measuring district compactness using graph theory. In *Annual Meeting of the Northeastern Political Science Association*.

DUCHIN, M. (2018). Outlier analysis for Pennsylvania congressional redistricting.

FIFIELD, B., HIGGINS, M., IMAI, K. and TARR, A. (2020a). Automated redistricting simulation using Markov chain Monte Carlo. *J. Comput. Graph. Statist.* **29** 715–728. MR4191238 https://doi.org/10.1080/10618600.2020.1739532

FIFIELD, B., IMAI, K., KAWAHARA, J. and KENNY, C. T. (2020b). The essential role of empirical validation in legislative redistricting simulation. *Stat. Public Policy* **7** 52–68.

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

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

GUTH, L., NIEH, A. and WEIGHILL, T. (2022). Three applications of entropy to gerrymandering. In *Political Geometry—Rethinking Redistricting in the US with Math, Law, and Everything in Between* 275–292. Birkhäuser/Springer, Cham. MR4436942 https://doi.org/10.1007/978-3-319-69161-9_14

HERSCHLAG, G., RAVIER, R. and MATTINGLY, J. C. (2017). Evaluating partisan gerrymandering in Wisconsin. arXiv preprint arXiv:1709.01596.

KENNY, C. T., MCCARTAN, C., FIFIELD, B. and IMAI, K. (2020). redist: Computational algorithms for redistricting simulation. https://CRAN.R-project.org/package=redist.

KENNY, C. T., MCCARTAN, C., SIMKO, T., KURIWAKI, S. and IMAI, K. (2023). Widespread partisan gerrymandering mostly cancels nationally, but reduces electoral competition. *Proc. Natl. Acad. Sci. USA* **120** e2217322120.

KOSTOCHKA, A. V. (1995). The number of spanning trees in graphs with a given degree sequence. *Random Structures Algorithms* **6** 269–274. MR1370962 https://doi.org/10.1002/rsa.3240060214

LEE, A. and WHITELEY, N. (2018). Variance estimation in the particle filter. *Biometrika* **105** 609–625. MR3842888 https://doi.org/10.1093/biomet/asy028

LEGLAND, F. and OUDJANE, N. (2005). A sequential particle algorithm that keeps the particle system alive. In 2005 13*th European Signal Processing Conference* 1–4. IEEE.

LIU, J. S., CHEN, R. and LOGVINENKO, T. (2001). A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo Methods in Practice. Stat. Eng. Inf. Sci.* 225–246. Springer, New York. MR1847794

LIU, J. S., CHEN, R. and WONG, W. H. (1998). Rejection control and sequential importance sampling. *J. Amer. Statist. Assoc.* **93** 1022–1031. MR1649197 https://doi.org/10.2307/2669846

LIU, Y. Y., CHO, W. K. T. and WANG, S. (2016). PEAR: A massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm Evol. Comput.* **30** 78–92.

MACMILLAN, W. (2001). Redistricting in a GIS environment: An optimisation algorithm using switching-points. *J. Geogr. Syst.* **3** 167–180.

MAGLEBY, D. B. and MOSESSON, D. B. (2018). A new approach for developing neutral redistricting plans. *Polit. Anal.* **26** 147–167.

MATTINGLY, J. C. and VAUGHN, C. (2014). Redistricting and the will of the people. arXiv preprint arXiv:1410.8796.

MCCARTAN, C. and IMAI, K. (2023). Supplement to "Sequential Monte Carlo for sampling balanced and compact redistricting plans." https://doi.org/10.1214/23-AOAS1763SUPPA, https://doi.org/10.1214/23-AOAS1763SUPPB

McCartan, C., Kenny, C. T., Simko, T., Garcia III, G., Wang, K., Wu, M., Kuriwaki, S. and Imai, K. (2022). Simulated redistricting plans for the analysis and evaluation of redistricting in the United States. *Sci. Data* **9** 689.

McKay, B. D. (1981). Spanning trees in random regular graphs. In *Proceedings of the Third Caribbean Conference on Combinatorics and Computing* (*Bridgetown*, 1981) 139–143. Univ. the West Indies, Cave Hill Campus, Barbados. MR0657198

Mehrotra, A., Johnson, E. L. and Nemhauser, G. L. (1998). An optimization based heuristic for political districting. *Manage. Sci.* **44** 1100–1114.

National Conference of State Legislatures (2021). Redistricting criteria. Available at https://www.ncsl.org/research/redistricting/redistricting-criteria.aspx.

League of Women Voters v. Commonwealth (2018). 178 A. 3d 737 (Pa: Supreme Court).

Olsson, J. and Douc, R. (2019). Numerically stable online estimation of variance in particle filters. *Bernoulli* **25** 1504–1535. MR3920380 https://doi.org/10.3150/18-bej1028

Peters, G. W., Fan, Y. and Sisson, S. A. (2012). On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. *Stat. Comput.* **22** 1209–1222. MR2992295 https://doi.org/10.1007/s11222-012-9315-y

Polsby, D. D. and Popper, R. D. (1991). The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law Policy Rev.* **9** 301–353.

Tutte, W. T. (1984). *Graph Theory*. *Encyclopedia of Mathematics and Its Applications* **21**. Addison-Wesley Company, Reading, MA. Advanced Book Program. With a foreword by C. St. J. A. Nash-Williams. MR0746795

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Anal.* **16** 667–718. Includes comments and discussions by seven discussants and a rejoinder by the authors. MR4298989 https://doi.org/10.1214/20-ba1221

Wilson, D. B. (1996). Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing* (*Philadelphia*, *PA*, 1996) 296–303. ACM, New York. MR1427525 https://doi.org/10.1145/237814.237880

Wu, L. C., Dou, J. X., Sleator, D., Frieze, A. and Miller, D. (2015). Impartial redistricting: A Markov Chain approach. arXiv preprint arXiv:1510.03247.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3324–3348
https://doi.org/10.1214/23-AOAS1764

ESTIMATING COVID-19 VACCINE PROTECTION RATES VIA DYNAMIC EPIDEMIOLOGICAL MODELS—A STUDY OF 10 COUNTRIES

BY YURU ZHU[1,a], JIA GU[1,b], YUMOU QIU[2,c] AND SONG XI CHEN[3,d]

[1]*Center for Statistical Science, Peking University,* [a]*yuruzhu@pku.edu.cn,* [b]*gujia@pku.edu.cn*
[2]*School of Mathematical Sciences and Center for Statistical Science, Peking University,* [c]*cqiuyumou@math.pku.edu.cn*
[3]*School of Mathematical Sciences and Guanghua School of Management, Peking University,* [d]*songxichen@pku.edu.cn*

The real-world performance of vaccines against COVID-19 infections is critically important to counter the pandemics. We propose a varying coefficient stochastic epidemic model to estimate the vaccine protection rates based on the publicly available epidemiological and vaccination data. To tackle the challenges posed by the unobserved state variables, we develop a multistep decentralized estimation procedure that uses different data segments to estimate different parameters. A B-spline structure is used to approximate the underlying infection rates and to facilitate model simulation in obtaining an objective function between the imputed and the simulation-based estimates of the latent state variables, leading to simulation-based estimation of the diagnosis rate using data in the prevaccine period and the vaccine effect parameters using data in the postvaccine periods. The time-varying infection, recovery and death rates are estimated by kernel regressions. We apply the proposed method to analyze the data in ten countries which collectively used eight vaccines. The analysis reveals that the average protection rate of the full vaccination was at least 22% higher than that of the partial vaccination and was largely above the WHO recognized level of 50% before November 20, 2021, including the Delta variant dominated period. The protection rates for the booster vaccine in the Omicron period were also provided.

REFERENCES

ALTARAWNEH, H. N., CHEMAITELLY, H., HASAN, M. R. et al. (2022). Protection against the omicron variant from previous SARS-CoV-2 infection. *N. Engl. J. Med.* **386** 1288–1290. PMID: 35139269. https://doi.org/10.1056/NEJMc2200133

ANDERSON, R. M. and MAY, R. M. (1982). Population dynamics of human helminth infections: Control by chemotherapy. *Nature* **297** 557–563. https://doi.org/10.1038/297557a0

AURANEN, K., ARJAS, E., LEINO, T. and TAKALA, A. K. (2000). Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. *J. Amer. Statist. Assoc.* **95** 1044–1053. MR1821713 https://doi.org/10.2307/2669741

BAAKE, E., BAAKE, M., BOCK, H. G. et al. (1992). Fitting ordinary differential equations to chaotic data. *Phys. Rev. A* **45** 5524–5529. https://doi.org/10.1103/PhysRevA.45.5524

BAROUCH, D. H. (2022). Covid-19 vaccines - immunity, variants, boosters. *N. Engl. J. Med.* **387** 1011–1020. https://doi.org/10.1056/NEJMra2206573

BERNAL, J. L., ANDREWS, N., GOWER, C. et al. (2021). Effectiveness of Covid-19 vaccines against the B.1.617.2 (Delta) variant. *N. Engl. J. Med.* **385** 585–594. https://doi.org/10.1056/NEJMoa2108891

BUITRAGO-GARCIA, D., EGLI-GANY, D., COUNOTTE, M. et al. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Med.* **17** e1003346. https://doi.org/10.1371/journal.pmed.1003346

CAI, C., LIU, Y., ZENG, S. et al. (2021a). The efficacy of COVID-19 vaccines against the B.1.617.2 (delta) variant. *Mol. Ther.* **29** 2890–2892. https://doi.org/10.1016/j.ymthe.2021.09.024

CAI, C., PENG, Y., SHEN, E. et al. (2021b). A comprehensive analysis of the efficacy and safety of COVID-19 vaccines. *Mol. Ther.* **29** 2794–2805. https://doi.org/10.1016/j.ymthe.2021.08.001

Key words and phrases. Reproduction number, scenario analysis, simulation-based estimation, stochastic epidemic model, varying coefficient model.

CDC (2022). Interim clinical considerations for use of COVID-19 vaccines currently approved or authorized in the United States. Available at https://www.cdc.gov/vaccines/covid-19/clinical-considerations/interim-considerations-us.html.

DASHTBALI, M. and MIRZAIE, M. (2021). A compartmental model that predicts the effect of social distancing and vaccination on controlling COVID-19. *Sci. Rep.* **11** 1–11.

DAVIS, C., LOGAN, N., TYSON, G., ORTON, R., HARVEY, W. T., PERKINS, J. S., MOLLETT, G., BLACOW, R. M., COVID-19 GENOMICS UK (COG-UK) CONSORTIUM et al. (2021). Reduced neutralisation of the Delta (B.1.617.2) SARS-CoV-2 variant of concern following vaccination. *PLoS Pathog.* **17** e1010022. https://doi.org/10.1371/journal.ppat.1010022

DONG, E., DU, H. and GARDNER, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20** 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

DORIA-ROSE, N., SUTHAR, M. S., MAKOWSKI, M. et al. (2021). Antibody persistence through 6 months after the second dose of mRNA-1273 vaccine for Covid-19. *N. Engl. J. Med.* **384** 2259–2261. https://doi.org/10.1056/NEJMc2103916

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

GIORDANO, G., COLANERI, M., DI FILIPPO, A. et al. (2021). Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat. Med.* **27** 993–998.

GUAN, W., NI, Z., HU, Y. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **382** 1708–1720. https://doi.org/10.1056/NEJMoa2002032

HAO, X., CHENG, S., WU, D. et al. (2020). Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584** 420–424.

HICKS, G. and RAY, W. (1971). Approximation methods for optimal control synthesis. *Can. J. Chem. Eng.* **49** 522–528.

HOLLAND, J. H. (1992). Genetic algorithms. *Sci. Amer.* **267** 66–73.

JOHNSON & JOHNSON (2021). Positive New Data for Johnson & Johnson Single-Shot COVID-19 Vaccine on Activity Against Delta Variant and Long-lasting Durability of Response. Available at https://www.jnj.com/positive-new-data-for-johnson-johnson-single-shot-covid-19-vaccine-on-activity-against-delta-variant-and-long-lasting-durability-of-response.

JONES, M. (1993). Simple boundary correction for kernel density estimation. *Stat. Comput.* **3** 135–146. https://doi.org/10.1007/BF00147776

KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A, Contain. Pap. Math. Phys. Character* **115** 700–721. https://doi.org/10.1098/rspa.1927.0118

KISLAYA, I., RODRIGUES, E. F., BORGES, V. et al. (2022). Comparative effectiveness of coronavirus vaccine in preventing breakthrough infections among vaccinated persons infected with Delta and Alpha variants. *Emerg. Infec. Dis.* **28** 331.

LEE, A. R. Y. B., WONG, S. Y., CHAI, L. Y. A. et al. (2022). Efficacy of Covid-19 vaccines in immunocompromised patients: Systematic review and meta-analysis. *BMJ* **376**.

LI, X., HUANG, Y., WANG, W. et al. (2021). Effectiveness of inactivated SARS-CoV-2 vaccines against the Delta variant infection in Guangzhou: A test-negative case–control real-world study. *Emerg. Microbes Infect.* **10** 1751–1759. https://doi.org/10.1080/22221751.2021.1969291

LIANG, H. and WU, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Amer. Statist. Assoc.* **103** 1570–1583. MR2504205 https://doi.org/10.1198/016214508000000797

LIU, C., GINN, H. M., DEJNIRATTISAI, W. et al. (2021). Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184** 4220–4236. https://doi.org/10.1016/j.cell.2021.06.020

MEDIĆ, S., ANASTASSOPOULOU, C., LOZANOV-CRVENKOVIĆ, Z. et al. (2022). Risk and severity of SARS-CoV-2 reinfections during 2020–2022 in Vojvodina, Serbia: A population-level observational study. *Lancet Reg. Health Eur.* **20** 100453. https://doi.org/10.1016/j.lanepe.2022.100453

PLANAS, D., VEYER, D., BAIDALIUK, A. et al. (2021). Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596** 276–280. https://doi.org/10.1038/s41586-021-03777-9

POLACK, F. P., THOMAS, S. J., KITCHIN, N. et al. (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.* **383** 2603–2615. https://doi.org/10.1056/NEJMoa2034577

QUICK, C., DEY, R. and LIN, X. (2021). Regression models for understanding COVID-19 epidemic dynamics with incomplete data. *J. Amer. Statist. Assoc.* **116** 1561–1577. MR4353694 https://doi.org/10.1080/01621459.2021.2001339

RAMSAY, J. O., HOOKER, G., CAMPBELL, D. and CAO, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 741–796. MR2368570 https://doi.org/10.1111/j.1467-9868.2007.00610.x

SHEIKH, A., MCMENAMIN, J., TAYLOR, B. et al. (2021). SARS-CoV-2 Delta VOC in Scotland: Demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397** 2461–2462. https://doi.org/10.1016/S0140-6736(21)01358-1

STATISTA (2021). Distribution of COVID-19 vaccine doses distributed to the European Economic Area (EEA) as of July 21, 2022, by manufacturer. Available at https://www.statista.com/statistics/1219343/covid19-vaccine-doses-distributed-in-europe-by-manufacturer.

THIRUVENGADAM, R., AWASTHI, A., MEDIGESHI, G. et al. (2021). Effectiveness of ChAdOx1 nCoV-19 vaccine against SARS-CoV-2 infection during the delta (B.1.617.2) variant surge in India: A test-negative, case-control study and a mechanistic study of post-vaccination immune responses. *Lancet Infect. Dis.* https://doi.org/10.1016/S1473-3099(21)00680-0

TIAN, T., TAN, J., LUO, W., JIANG, Y., CHEN, M., YANG, S., WEN, C., PAN, W. and WANG, X. (2021). The effects of stringent and mild interventions for coronavirus pandemic. *J. Amer. Statist. Assoc.* **116** 481–491. MR4269997 https://doi.org/10.1080/01621459.2021.1897015

TOWNSEND, J. P., HASSLER, H. B., WANG, Z. et al. (2021). The durability of immunity against reinfection by SARS-CoV-2: A comparative evolutionary study. *Lancet Microbe* **2** e666–e675. https://doi.org/10.1016/S2666-5247(21)00219-6

VOYSEY, M., CLEMENS, S. A. C., MADHI, S. A. et al. (2021). Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: An interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* **397** 99–111. https://doi.org/10.1016/S0140-6736(20)32661-1

WHO (2021). What is COVID-19 vaccine efficacy. Available at https://www.afro.who.int/news/what-covid-19-vaccine-efficacy.

YAN, H., ZHU, Y., GU, J., HUANG, Y., SUN, H., ZHANG, X., WANG, Y., QIU, Y. and CHEN, S. X. (2021). Better strategies for containing COVID-19 pandemic: A study of 25 countries via a vSIADR model. *Proc. R. Soc. A* **477** 20200440. MR4258333

ZHU, Y., GU, J., QIU, Y. and CHEN, S. X. (2023). Supplement to "Estimating COVID-19 vaccine protection rates via dynamic epidemiological models—a study of 10 countries." https://doi.org/10.1214/23-AOAS1764SUPP

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3349–3362
https://doi.org/10.1214/23-AOAS1765

ESTIMATING COVID-19 TRANSMISSION TIME USING HAWKES POINT PROCESSES

BY FREDERIC SCHOENBERG[a]

Department of Statistics, University of California, Los Angeles, [a]*frederic@stat.ucla.edu*

The question addressed here is whether, using Hawkes models. the distribution of SARS-CoV-2 (Covid-19) transmission times can be estimated accurately with only case-count data. We fit Hawkes models with varying productivities to each of the 50 United States individually, estimating for each state a transmission time density, both nonparametrically and using a normal approximation. We find that, for nearly all states, the estimated transmission times are centered near seven days with a standard deviation of approximately one day. Compared to previous reports, the results here suggest that transmission times for SARS-CoV-2 are somewhat shorter, on average, and the distribution is less diffuse, though the results also suggest the possibility of transmission occurring on the first day of exposure.

REFERENCES

BAJEMA, K. L., WIEGAND, R. E., CUFFE, K., PATEL, S. V., IACHAN, R., LIM, T., LEE, A., MOYSE, D., HAVERS, F. et al. (2021). Estimated SARS-CoV-2 seroprevalence in the US as of September 2020. *JAMA Intern. Med.* **181** 450–460.

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

BERTSIMAS, D. (2020). MIT Covidanalytics, May 2020. Available at covidanalytics.io [Online; accessed 24-May-2020].

BRAY, A., WONG, K., BARR, C. D. and SCHOENBERG, F. P. (2014). Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *Ann. Appl. Stat.* **8** 2247–2267. MR3292496 https://doi.org/10.1214/14-AOAS767

BRILLINGER, D. R. (1981). *Time Series*: *Data Analysis and Theory*, 2nd ed. *Holden-Day Series in Time Series Analysis*. Holden-Day, Inc., Oakland, CA. MR0595684

CAUCHEMEZ, S., BOELLE, P. Y., DONNELLY, C. A., FERGUSON, N. M., THOMAS, G., LEUNG, G. M., HEDLEY, A. J., ANDERSON, R. M. and VALLERON, A. J. (2006). Real-time estimates in early detection of SARS. *Emerg. Infec. Dis.* **12** 110.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021a). Available at https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html, last accessed 9/14/21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021b). Available at https://www.cdc.gov/coronavirus/2019-ncov/your-health/quarantine-isolation.html, last accessed 9/14/21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021c). Available at https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html, last accessed 9/14/21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021d). Available at https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html, last accessed 9/14/21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021e). Available at https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm, last accessed 9/14/21.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC) (2021f). Available at https://covid.cdc.gov/covid-data-tracker, last accessed 9/14/21.

CHIANG, W. H., LIU, X. and MOHLER, G. (2020). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *medRxiv*. https://doi.org/10.1101/2020.06.06.20124149.

CLEMENTS, R. A., SCHOENBERG, F. P. and SCHORLEMMER, D. (2011). Residual analysis methods for space-time point processes with applications to earthquake forecast models in California. *Ann. Appl. Stat.* **5** 2549–2571. MR2907126 https://doi.org/10.1214/11-AOAS487

Key words and phrases. Disease epidemics, Hawkes model, point process, self-exciting.

CLEMENTS, R. A., SCHOENBERG, F. P. and VEEN, A. (2012). Evaluation of space-time point process models using super-thinning. *Environmetrics* **23** 606–616. MR3020078 https://doi.org/10.1002/env.2168

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes*: *Elementary Theory and Methods*. *Vol*. *I*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

DYE, C. and GAY, N. (2003). Modeling the SARS epidemic. *Science* **300** 1884–1885.

FARRINGTON, C., KANAAN, M. and GAY, N. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* **4** 279–295.

GORDON, J. S., CLEMENTS, R. A., SCHOENBERG, F. P. and SCHORLEMMER, D. (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *Spat*. *Stat*. **14** 133–150. MR3429717 https://doi.org/10.1016/j.spasta.2015.06.001

GUAN, W. J., NI, Z. Y., HU, Y. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *N*. *Engl*. *J*. *Med*. **382** 1708–1720.

HAWKES, A. G. (1971). Point spectra of some mutually exciting point processes. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **33** 438–443. MR0358976

HUANG, C., WANG, Y., LI, X., REN, L., ZHAO, J., HU, Y., ZHANG, L., FAN, G., XU, J. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395** 497–506.

INSTITUTE FOR HEALTH METRICS AND EVALUATION (IHME) (2020). IHME Covid-19 predictions, May 2020. Available at covid19.healthdata.org [Online; accessed 24-May-2020].

JEWELL, N. P., LEWNARD, J. A. and JEWELL, B. L. (2020). Caution warranted: Using the institute for health metrics and evaluation model for predicting the course of the Covid-19 pandemic. *Ann*. *Intern*. *Med*. **173** 226–227.

KELLY, J. D., HARRIGAN, R. J., PARK, J., HOFF, N. A., LEE, S. D., WANNIER, R., SELO, B., MOSSOKO, M., NJOKOLO, B. et al. (2019). Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of Congo using Hawkes point process models. *Epidemics* **28** 100354.

KIRCHNER, M. (2016). Hawkes and INAR(∞) processes. *Stochastic Process*. *Appl*. **126** 2494–2525. MR3505235 https://doi.org/10.1016/j.spa.2016.02.008

KIRCHNER, M. (2017). An estimation procedure for the Hawkes process. *Quant*. *Finance* **17** 571–595. MR3620953 https://doi.org/10.1080/14697688.2016.1211312

KRESIN, C., SCHOENBERG, F. and MOHLER, G. (2021). Comparison of Hawkes and SEIR models for the spread of Covid-19. *Adv*. *Appl*. *Stat*. **74** 83–106.

LAUER, S. A., GRANTZ, K. H., BI, Q., JONES, F. K., ZHENG, Q., MEREDITH, H. R., AZMAN, A. S., REICH, N. G. and LESSLER, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann*. *Intern*. *Med*. **172** 577–582.

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

LI, Q., GUAN, X., WU, P., WANG, X., ZHOU, L., TONG, Y., REN, R., LEUNG, K. S. M., LAU, E. H. Y. et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N*. *Engl*. *J*. *Med*. **382** 1199–1207.

LOS ALAMOS NATIONAL LABORATORY (LANL) (2020). Covid-19 confirmed and forecasted case data, May 2020. Available at covid-19.bsvgateway.org [Online; accessed 24-May-2020].

LOTFI, M., HAMBLIN, M. R. and REZAEI, M. (2020). COVID-19: Transmission, prevention, and potential therapeutic opportunities. *Clin*. *Chim*. *Acta* **508** 254–266.

MARSAN, D. and LENGLINÉ, O. (2008). Extending earthquakes' reach through cascading. *Science* **319** 1076–1079.

MEYER, S., ELIAS, J. and HÖHLE, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics* **68** 607–616. MR2959628 https://doi.org/10.1111/j.1541-0420.2011.01684.x

MEYERS, L. A. (2007). Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull*. *Amer*. *Math*. *Soc*. (*N*.*S*.) **44** 63–86. MR2265010 https://doi.org/10.1090/S0273-0979-06-01148-7

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

MOHLER, G., SCHOENBERG, F., SHORT, M. B. and SLEDGE, D. (2021). Analyzing the impacts of public policy on COVID-19 transmission: A case study of the role of model and dataset selection using data from Indiana. *Stat*. *Public Policy* **8** 1–8.

OGATA, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann*. *Inst*. *Statist*. *Math*. **30** 243–261. MR0514494 https://doi.org/10.1007/BF02480216

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

PARK, J., CHAFFEE, A. W., HARRIGAN, R. J. and SCHOENBERG, F. P. (2020). A non-parametric Hawkes model of the spread of Ebola in West Africa. *J. Appl. Stat.* **49** 621–637. MR4381539 https://doi.org/10.1080/02664763.2020.1825646

PARK, J., SCHOENBERG, F. P., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2021). Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates. *J. Amer. Statist. Assoc.* **116** 1674–1687. MR4353705 https://doi.org/10.1080/01621459.2021.1898408

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

REINHART, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33** 299–318. MR3843374 https://doi.org/10.1214/17-STS629

RIZOIU, M. A., MISHRA, S., KONG, Q., CARMAN, M. and XIE, L. (2018). Sir-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In *Proceedings of the* 2018 *World Wide Web Conference* 419–428.

SCHOENBERG, F. (2023). Supplement to "Estimating Covid-19 transmission time using Hawkes point processes." https://doi.org/10.1214/23-AOAS1765SUPP

SCHOENBERG, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statist. Sinica* **26** 861–879. MR3497774

SCHOENBERG, F. P. (2022). Nonparametric estimation of variable productivity Hawkes processes. *Environmetrics* **33** Paper No. e2747, 13. MR4476429

SCHOENBERG, F. P., HOFFMANN, M. and HARRIGAN, R. J. (2019). A recursive point process model for infectious diseases. *Ann. Inst. Statist. Math.* **71** 1271–1287. MR3993533 https://doi.org/10.1007/s10463-018-0690-9

SCHORLEMMER, D., WERNER, M. J., MARZOCCHI, W., JORDAN, T. H., OGATA, Y., JACKSON, D. D., MAK, S., RHOADES, D. A., GERSTENBERGER, M. C. et al. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismol. Res. Lett.* **89** 1305–1313.

WALLINGA, J. and TEUNIS, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Amer. J. Epidemiol.* **160** 509–516.

WANG, D., HU, B., HU, C., ZHU, F., LIU, X., ZHANG, J., WANG, B., XIANG, H., CHENG, Z. et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323** 1061–1069. https://doi.org/10.1001/jama.2020.1585

WERMER, E. and STEIN, J. (2020). Trump administration pushing to block new money for testing, tracing and CDC in upcoming coronavirus relief bill. *Washington Post*, 07/18/20. Available at https://www.washingtonpost.com/us-policy/2020/07/18/white-house-testing-budget-cdc-coronavirus.

WORLD HEALTH ORGANIZATION (WHO) (2021). Available at https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions, last accessed 9/14/21.

XU, R. H., HE, J. F., EVANS, M. R., PENG, G. W., FIELD, H. E., YU, D. W., LEE, C. K., LUO, H. M., LIN, W. S. et al. (2004). Epidemiological clues to SARS origin in. *China Emerg. Infect. Dis.* **10** 1030–1037.

YANG, A. S. (2019). Modeling the transmission dynamics of pertussis using recursive point process and SEIR model. Ph.D. thesis, UCLA.

YANG, X., YU, Y., XU, J., SHU, H., XIA, J., LIU, H., WU, Y., ZHANG, L., YU, Z. et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study. *Lancet Respir. Med.* **8** 475–481.

YOU, C., DENG, Y., HU, W., SUN, J., LIN, Q., ZHOU, F., PANG, C. H., ZHANG, Y., CHEN, Z. et al. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *Int. J. Hyg. Environ. Health* 113555.

YUAN, B., SCHOENBERG, F. P. and BERTOZZI, A. L. (2021). Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction. *Ann. Inst. Statist. Math.* **73** 1127–1152. MR4330314 https://doi.org/10.1007/s10463-020-00780-1

ZECHAR, J. D., SCHORLEMMER, D., WERNER, M. J., GERSTENBERGER, M. C., RHOADES, D. A. and JORDAN, T. H. (2013). Regional earthquake likelihood models I: First-order results. *Bull. Seismol. Soc. Amer.* **103** 787–798.

ZHOU, F., YU, T., DU, R., FAN, G., LIU, Y., LIU, Z., XIANG, J., WANG, Y., SONG, B. et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395** 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res., Solid Earth* **109**.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3363–3383
https://doi.org/10.1214/23-AOAS1766

JOINT STOCHASTIC SIMULATION OF EXTREME COASTAL AND OFFSHORE SIGNIFICANT WAVE HEIGHTS

BY JULIETTE LEGRAND[1,a], PIERRE AILLIOT[2,c], PHILIPPE NAVEAU[1,b] AND NICOLAS RAILLARD[3,d]

[1]*Laboratoire des Sciences du Climat et de l'Environnement, UMR8212 CEA-CNRS-UVSQ, IPSL & Université Paris-Saclay,* [a]*juliette.legrand1@inrae.fr,* [b]*philippe.naveau@lsce.ipsl.fr*

[2]*Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Occidentale,* [c]*pierre.ailliot@univ-brest.fr*

[3]*IFREMER, RDT,* [d]*Nicolas.Raillard@ifremer.fr*

The characterisation of future extreme wave events is crucial because of their multiple impacts, covering a broad range of topics such as coastal flood hazard, coastal erosion, reliability of offshore and coastal structures. The main goal of this paper is to propose and study a stochastic simulator that, given offshore conditions (peak direction D_p, peak period T_p and moderately high significant wave heights H_s), produces jointly offshore and coastal extreme H_s, a quantity measuring the wave severity and which represent a key feature in coastal risk analysis. For this purpose we rely on bivariate Peaks over Threshold, and a nonparametric simulation scheme of bivariate GPD is developed. From this joint simulator, a second generator is derived, allowing for conditional simulations of extreme H_s. Finally, to take into account nonstationarities, the extended generalised Pareto model is also adapted, letting the parameters vary with specific sea-state parameters T_p and D_p. The performances of the two proposed generators are illustrated on simulated data and then applied to the simulation of new extreme oceanographic conditions close to the French Brittany coast using hindcast sea-state data. Results show that the proposed algorithms successfully simulate future extreme H_s near the coast in a nonparametric way, jointly or conditionally on sea-state parameters from a coarser model.

REFERENCES

ACCENSI, M. and MAISONDIEU, C. (2015). HOMERE. Available at https://doi.org/10.12770/cf47e08d-1455-4254-955e-d66225c9dc90.

ARDHUIN, F. and ACCENSI, M. (2014). IOWAGA. Available at https://sextant.ifremer.fr/record/c87f6f24-63b4-46ec-b40e-f185a61dc672/.

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes*: *Theory and Applications*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2108013 https://doi.org/10.1002/0470012382

BERTIN, X., BRUNEAU, N., BREILH, J.-F., FORTUNATO, A. B. and KARPYTCHEV, M. (2012). Importance of wave age and resonance in storm surges: The case Xynthia, Bay of Biscay. *Ocean Model*. **42** 16–30.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann*. *Statist*. **9** 1196–1217. MR0630103

CAIRES, S. and STERL, A. (2005). 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *J*. *Climate* **18** 1032–1048.

CASAS-PRAT, M., WANG, X. L. and SIERRA, J. P. (2014). A physical-based statistical method for modeling ocean wave heights. *Ocean Model*. **73** 59–75.

CASTILLO, E. and SARABIA, J. M. (1992). Engineering analysis of extreme value data: Selection of models. *Journal of Waterway*, *Port*, *Coastal*, *and Ocean Engineering* **118** 129–146.

CHAVEZ-DEMOULIN, V. and DAVISON, A. C. (2005). Generalized additive modelling of sample extremes. *J*. *R*. *Stat*. *Soc*., *Ser*. *C* **54** 207–222. MR2134607 https://doi.org/10.1111/j.1467-9876.2005.00479.x

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J*. *Amer*. *Statist*. *Assoc*. **83** 596–610.

Key words and phrases. Bivariate extremes, multivariate generalised Pareto distribution, simulation of extremes, nonstationarity, extended generalised Pareto distribution, covariate effects, significant wave heights.

COLE, T. J. and GREEN, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat. Med.* **11** 1305–1319. https://doi.org/10.1002/sim.4780111005
COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics*. Springer London, Ltd., London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0
COLES, S., HEFFERNAN, J. E. and TAWN, J. A. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.
COLLINS, M., SUTHERLAND, M., BOUWER, L., CHEONG, S. M., FRÖLICHER, T., COMBES, H. J. D., ROXY, M. K., LOSADA, I., MCINNES, K. et al. (2019). Extremes, abrupt changes and managing risk. In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* (H. O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai et al., eds.) Cambridge Univ. Press, Cambridge.
DE CARVALHO, M., PEREIRA, S., PEREIRA, P. and DE ZEA BERMUDEZ, P. (2022). An extreme value Bayesian Lasso for the conditional left and right tails. *J. Agric. Biol. Environ. Stat.* **27** 222–239. MR4416781 https://doi.org/10.1007/s13253-021-00469-9
DE LEO, F., BESIO, G., BRIGANTI, R. and VANEM, E. (2021). Non-stationary extreme value analysis of sea states based on linear trends. Analysis of annual maxima series of significant wave height and peak period in the Mediterranean Sea. *Coastal Eng.* **167** 103896.
EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
EWANS, K. and JONATHAN, P. (2008). The effect of directionality on northern North Sea extreme wave design criteria. *Journal of Offshore Mechanics and Arctic Engineering* **130**.
FELD, G., RANDELL, D., WU, Y., EWANS, K. and JONATHAN, P. (2014). Estimation of storm peak and intra-storm directional-seasonal design conditions in the North Sea. *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering—OMAE* **4**.
GENOVESE, E. and PRZYLUSKI, V. (2013). Storm surge disaster risk management: The Xynthia case study in France. *Journal of Risk Research* **16** 825–841.
GUMBEL, E. J. (1961). Bivariate logistic distributions. *J. Amer. Statist. Assoc.* **56** 335–349. MR0158451
HARUNA, A., BLANCHET, J. and FAVRE, A. C. (2022). Performance-based comparison of regionalization methods to improve the at-site estimates of daily precipitation. *Hydrol. Earth Syst. Sci.* **26** 2797–2811.
HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 497–546. MR2088289 https://doi.org/10.1111/j.1467-9868.2004.02050.x
HEREDIA-ZAVONI, E. and MONTES-ITURRIZAGA, R. (2019). Modeling directional environmental contours using three dimensional vine copulas. *Ocean Eng.* **187** 106102.
HOLTHUIJSEN, L. H. (2007). *Waves in Oceanic and Coastal Waters*. Cambridge University Press, Cambridge.
JONATHAN, P. and EWANS, K. (2007). The effect of directionality on extreme wave design criteria. *Ocean Eng.* **34** 1977–1994.
JONATHAN, P. and EWANS, K. (2013). Statistical modelling of extreme ocean environments for marine design: A review. *Ocean Eng.* **62** 91–109.
JONATHAN, P., EWANS, K. and RANDELL, D. (2014). Non-stationary conditional extremes of northern North Sea storm characteristics. *Environmetrics* **25** 172–188. MR3200308 https://doi.org/10.1002/env.2262
KIRILIOUK, A., ROOTZÉN, H., SEGERS, J. and WADSWORTH, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics* **61** 123–135. MR3933664 https://doi.org/10.1080/00401706.2018.1462738
KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417 https://doi.org/10.1093/biomet/81.4.673
LE CARRER, N. (2022). egpd4gamlss. Available at https://github.com/noemielc/egpd4gamlss.
LEGRAND, J., AILLIOT, P., NAVEAU, P. and RAILLARD, N. (2023). Supplement to "Joint stochastic simulation of extreme coastal and offshore significant wave heights." https://doi.org/10.1214/23-AOAS1766SUPPA, https://doi.org/10.1214/23-AOAS1766SUPPB
MARCON, G., NAVEAU, P. and PADOAN, S. (2017). A semi-parametric stochastic generator for bivariate extreme events. *Stat* **6** 184–201. MR3653052 https://doi.org/10.1002/sta4.145
MARSHALL, A. W. and OLKIN, I. (1967). A multivariate exponential distribution. *J. Amer. Statist. Assoc.* **62** 30–44. MR0215400
MÉNDEZ, F. J., MENÉNDEZ, M., LUCEÑO, A., MEDINA, R. and GRAHAM, N. E. (2008). Seasonality and duration in extreme value distributions of significant wave height. *Ocean Eng.* **35** 131–138.
MICHEL, R. (2006). Simulation and estimation in multivariate generalized Pareto models. Ph.D. thesis, Univ. Würzburg.
MOUSLIM, H., BABARIT, A., CLÉMENT, A. and BORGARINO, B. (2009). Development of the French wave energy test site SEM-REV. In *Proceedings of the 8th European Wave and Tidal Energy Conference* 31–35.
NAVEAU, P., HUSER, R., RIBEREAU, P. and HANNART, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resour. Res.* **52** 2753–2769.

NICOLAE LERMA, A., BULTEAU, T., LECACHEUX, S. and IDIER, D. (2015). Spatial variability of extreme wave height along the Atlantic and channel French coast. *Ocean Eng*. **97** 175–185.

NORTHROP, P. J. and JONATHAN, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics* **22** 799–809. MR2861046 https://doi.org/10.1002/env.1106

PAPASTATHOPOULOS, I. and TAWN, J. A. (2013). Extended generalised Pareto models for tail estimation. *J. Statist. Plann. Inference* **143** 131–143. MR2969016 https://doi.org/10.1016/j.jspi.2012.07.001

RIVOIRE, P., MARTIUS, O. and NAVEAU, P. (2021). A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets. *Earth Space Sci*. **8** e2020EA001633.

ROOTZÉN, H., SEGERS, J. and WADSWORTH, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *J. Multivariate Anal*. **165** 117–131. MR3768756 https://doi.org/10.1016/j.jmva.2017.12.003

ROOTZÉN, H. and TAJVIDI, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli* **12** 917–930. MR2265668 https://doi.org/10.3150/bj/1161614952

ROSS, E., RANDELL, D., EWANS, K., FELD, G. and JONATHAN, P. (2017). Efficient estimation of return value distributions from non-stationary marginal extreme value models using Bayesian inference. *Ocean Eng*. **142** 315–328.

SENEVIRATNE, S. I., ZHANG, X., ADNAN, M., BADI, D. C., DI LUCA, A., GHOSH, S., ISKANDAR, I., KOSSIN, J., LEWIS, S., OTTO, F., PINTO, I., SATOH, M., VICENTE-SERRANO, S. M. and WEHNER, M. (2021). Weather and climate extreme events in a changing climate. In *Climate Change* 2021: *The Physical Science Basis*. *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou, eds.) Cambridge University Press, Cambridge.

SHOOTER, R., ROSS, E., RIBAL, A., YOUNG, I. R. and JONATHAN, P. (2021). Spatial dependence of extreme seas in the North East Atlantic from satellite altimeter measurements. *Environmetrics* **32** Paper No. e2674. MR4259540 https://doi.org/10.1002/env.2674

SHOOTER, R., ROSS, E., RIBAL, A., YOUNG, I. R. and JONATHAN, P. (2022). Multivariate spatial conditional extremes for extreme ocean environments. *Ocean Eng*. **247** 110647.

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

SHOOTER, R., TAWN, J. A., ROSS, E. and JONATHAN, P. (2021). Basin-wide spatial conditional extremes for severe ocean storms. *Extremes* **24** 241–265. MR4246277 https://doi.org/10.1007/s10687-020-00389-w

STASINOPOULOS, M., RIGBY, B. and AKANTZILIOTOU, C. (2008). *Instructions on How to Use the Gamlss Package in R Second Edition*.

TENCALIEC, P., FAVRE, A.-C., NAVEAU, P., PRIEUR, C. and NICOLET, G. (2019). Flexible semiparametric generalized Pareto modeling of the entire range of rainfall amount. *Environmetrics* **31** e2582. MR4075861 https://doi.org/10.1002/env.2582

TENDIJCK, S., EASTOE, E. F., TAWN, J. A., RANDELL, D. and JONATHAN, P. (2021). Modeling the extremes of bivariate mixture distributions with application to oceanographic data. *J. Amer. Statist. Assoc*. **0** 1–12.

TOWE, R., EASTOE, E. F., TAWN, J. A. and JONATHAN, P. (2017). Statistical downscaling for future extreme wave heights in the North Sea. *Ann. Appl. Stat*. **11** 2375–2403. MR3743301 https://doi.org/10.1214/17-AOAS1084

VANEM, E. and FAZERES-FERRADOSA, T. (2022). A truncated, translated Weibull distribution for shallow water sea states. *Coastal Eng*. **172** 104077.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3384–3402
https://doi.org/10.1214/23-AOAS1767

A RELUCTANT ADDITIVE MODEL FRAMEWORK FOR INTERPRETABLE NONLINEAR INDIVIDUALIZED TREATMENT RULES

BY JACOB M. MARONGE[1,a], JARED D. HULING[2,b] AND GUANHUA CHEN[3,c]

[1]*Department of Biostatistics, University of Texas MD Anderson Cancer Center,* [a]*jmmaronge@gmail.com*
[2]*Division of Biostatistics, University of Minnesota,* [b]*huling@umn.edu*
[3]*Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison,* [c]*gchen25@wisc.edu*

Individualized treatment rules (ITRs) for treatment recommendation is an important topic for precision medicine as not all beneficial treatments work well for all individuals. Interpretability is a desirable property of ITRs, as it helps practitioners make sense of treatment decisions, yet there is a need for ITRs to be flexible to effectively model complex biomedical data for treatment decision making. Many ITR approaches either focus on linear ITRs, which may perform poorly when true optimal ITRs are nonlinear, or black-box nonlinear ITRs, which may be hard to interpret and can be overly complex. This dilemma indicates a tension between interpretability and accuracy of treatment decisions. Here we propose an additive model-based nonlinear ITR learning method that balances interpretability and flexibility of the ITR. Our approach aims to strike this balance by allowing both linear and nonlinear terms of the covariates in the final ITR. Our approach is parsimonious in that the nonlinear term is included in the final ITR only when it substantially improves the ITR performance. To prevent overfitting, we combine crossfitting and a specialized information criterion for model selection. Through extensive simulations we show that our methods are data-adaptive to the degree of nonlinearity and can favorably balance ITR interpretability and flexibility. We further demonstrate the robust performance of our methods with an application to a cancer drug sensitive study.

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

BIAN, Z., MOODIE, E. E. M., SHORTREED, S. M. and BHATNAGAR, S. (2023). Variable selection in regression-based estimation of dynamic treatment regimes. *Biometrics* **79** 988–999. MR4606331 https://doi.org/10.1111/biom.13608

CHEN, G., ZENG, D. and KOSOROK, M. R. (2016). Personalized dose finding using outcome weighted learning. *J. Amer. Statist. Assoc.* **111** 1509–1521. MR3601705 https://doi.org/10.1080/01621459.2016.1148611

CHEN, S., TIAN, L., CAI, T. and YU, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209. MR3744534 https://doi.org/10.1111/biom.12676

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

FAN, C., LU, W., SONG, R. and ZHOU, Y. (2017). Concordance-assisted learning for estimating optimal individualized treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1565–1582. MR3731676 https://doi.org/10.1111/rssb.12216

GHANDI, M., HUANG, F. W., JANÉ-VALBUENA, J., KRYUKOV, G. V., LO, C. C., MCDONALD, E. R., BARRETINA, J., GELFAND, E. T., BIELSKI, C. M. et al. (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569** 503–508.

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

Key words and phrases. Individualized treatment rules, precision medicine, reluctant additive models.

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

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

HULING, J. D. and YU, M. (2021). Subgroup identification using the personalized package. *J. Stat. Softw.* **98** 1–60.

IORIO, F., KNIJNENBURG, T. A., VIS, D. J., BIGNELL, G. R., MENDEN, M. P., SCHUBERT, M. et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* **166** 740–754.

KRIVOBOKOVA, T., CRAINICEANU, C. M. and KAUERMANN, G. (2008). Fast adaptive penalized splines. *J. Comput. Graph. Statist.* **17** 1–20. MR2424792 https://doi.org/10.1198/106186008X287328

LIANG, M., YE, T. and FU, H. (2018). Estimating individualized optimal combination therapies through outcome weighted deep learning algorithms. *Stat. Med.* **37** 3869–3886. MR3873688 https://doi.org/10.1002/sim.7902

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

MARONGE, J. M., HULING, J. D. and CHEN, G. (2023). Supplement to "A reluctant additive model framework for interpretable nonlinear individualized treatment rules." https://doi.org/10.1214/23-AOAS1767SUPPA, https://doi.org/10.1214/23-AOAS1767SUPPB

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

MI, X., ZOU, F. and ZHU, R. (2019). Bagging and deep learning in optimal individualized treatment rules. *Biometrics* **75** 674–684. MR3999189 https://doi.org/10.1111/biom.12990

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

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

PAN, Y. and ZHAO, Y.-Q. (2021). Improved doubly robust estimation in learning optimal individualized treatment rules. *J. Amer. Statist. Assoc.* **116** 283–294. MR4227694 https://doi.org/10.1080/01621459.2020.1725522

PARK, H., PETKOVA, E., TARPEY, T. and OGDEN, R. T. (2022). A sparse additive model for treatment effect-modifier selection. *Biostatistics* **23** 412–429. MR4424978 https://doi.org/10.1093/biostatistics/kxaa032

PETERSEN, A. and WITTEN, D. (2019). Data-adaptive additive modeling. *Stat. Med.* **38** 583–600. MR3902599 https://doi.org/10.1002/sim.7859

PETERSEN, A., WITTEN, D. and SIMON, N. (2016). Fused lasso additive model. *J. Comput. Graph. Statist.* **25** 1005–1025. MR3572026 https://doi.org/10.1080/10618600.2015.1073155

QI, Z., LIU, D., FU, H. and LIU, Y. (2020). Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes. *J. Amer. Statist. Assoc.* **115** 678–691. MR4107672 https://doi.org/10.1080/01621459.2018.1529597

QIAN, M. and MURPHY, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210. MR2816351 https://doi.org/10.1214/10-AOS864

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

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071 https://doi.org/10.1198/016214504000001880

RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1** 206.

SHI, C., SONG, R. and LU, W. (2021). Concordance and value information criteria for optimal treatment decision. *Ann. Statist.* **49** 49–75. MR4206669 https://doi.org/10.1214/19-AOS1908

TAY, J. K. and TIBSHIRANI, R. (2020). Reluctant generalized additive modeling. *Int. Stat. Rev.* **88** S205–S224.

TIAN, L., ALIZADEH, A. A., GENTLES, A. J. and TIBSHIRANI, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Amer. Statist. Assoc.* **109** 1517–1532. MR3293607 https://doi.org/10.1080/01621459.2014.951443

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

VAYENA, E., BLASIMME, A. and COHEN, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **15** e1002689.

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

WAHBA, G. (2006). *Splines in Nonparametric Regression* **4**. Wiley Online Library, London, UK.

WALLACE, M. P. and MOODIE, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71** 636–644. MR3402599 https://doi.org/10.1111/biom.12306

YANG, W., SOARES, J., GRENINGER, P., EDELMAN, E. J., LIGHTFOOT, H., FORBES, S., BINDAL, N., BEARE, D., SMITH, J. A. et al. (2012). Genomics of drug sensitivity in cancer (gdsc): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. **41** D955–D961.

YU, G., BIEN, J. and TIBSHIRANI, R. (2019). Reluctant interaction modeling. Preprint. Available at: arXiv:1907.08414.

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

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc*. **112** 169–187. MR3646564 https://doi.org/10.1080/01621459.2015.1093947

ZHU, R., ZHAO, Y.-Q., CHEN, G., MA, S. and ZHAO, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics* **73** 391–400. MR3665956 https://doi.org/10.1111/biom.12593

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3403–3425
https://doi.org/10.1214/23-AOAS1768

MULTIMODEL ENSEMBLE ANALYSIS WITH NEURAL NETWORK GAUSSIAN PROCESSES

BY TREVOR HARRIS[1,a], BO LI[2,b] AND RYAN SRIVER[3,c]

[1]*Department of Statistics, Texas A&M University,* [a]*tharris@tamu.edu*
[2]*Department of Statistics, University of Illinois at Urbana Champaign,* [b]*libo@illinois.edu*
[3]*Department of Atmospheric Sciences, University of Illinois at Urbana Champaign,* [c]*rsriver@illinois.edu*

Multimodel ensemble analysis integrates information from multiple climate models into a unified projection. However, existing integration approaches, based on model averaging, can dilute fine-scale spatial information and incur bias from rescaling low-resolution climate models. We propose a statistical approach, called NN-GPR, using Gaussian process regression (GPR) with an infinitely wide deep neural network based covariance function. NN-GPR requires no assumptions about the relationships between climate models, no interpolation to a common grid, and automatically downscales as part of its prediction algorithm. Model experiments show that NN-GPR can be highly skillful at surface temperature and precipitation forecasting by preserving geospatial signals at multiple scales and capturing interannual variability. Our projections particularly show improved accuracy and uncertainty quantification skill in regions of high variability, which allows us to cheaply assess tail behavior at a 0.44°/50 km spatial resolution without a regional climate model (RCM). Evaluations on reanalysis data and SSP2-4.5 forced climate models show that NN-GPR produces similar, overall climatologies to the model ensemble while better capturing fine-scale spatial patterns. Finally, we compare NN-GPR's regional predictions against two RCMs and show that NN-GPR can rival the performance of RCMs using only global model data as input.

REFERENCES

ABRAMOWITZ, G., HERGER, N., GUTMANN, E., HAMMERLING, D., KNUTTI, R., LEDUC, M., LORENZ, R., PINCUS, R. and SCHMIDT, G. A. (2019). ESD reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.* **10** 91–105.

ALEMOHAMMAD, S., WANG, Z., BALESTRIERO, R. and BARANIUK, R. (2020). The recurrent neural tangent kernel. Preprint. Available at arXiv:2006.10246.

ARORA, S., DU, S. S., LI, Z., SALAKHUTDINOV, R., WANG, R. and YU, D. (2019). Harnessing the power of infinitely wide deep nets on small-data tasks. Preprint. Available at arXiv:1910.01663.

BATTAGLIA, P. W., HAMRICK, J. B., BAPST, V., SANCHEZ-GONZALEZ, A., ZAMBALDI, V., MALINOWSKI, M., TACCHETTI, A., RAPOSO, D., SANTORO, A. et al. (2018). Relational inductive biases, deep learning, and graph networks. Preprint. Available at arXiv:1806.01261.

BHAT, K. S., HARAN, M., TERANDO, A. and KELLER, K. (2011). Climate projections using Bayesian model averaging and space-time dependence. *J. Agric. Biol. Environ. Stat.* **16** 606–628. MR2862301 https://doi.org/10.1007/s13253-011-0069-3

BORNN, L., SHADDICK, G. and ZIDEK, J. V. (2012). Modeling nonstationary processes through dimension expansion. *J. Amer. Statist. Assoc.* **107** 281–289. MR2949359 https://doi.org/10.1080/01621459.2011.646919

BOWMAN, K. W., CRESSIE, N., QU, X. and HALL, A. (2018). A hierarchical statistical framework for emergent constraints: Application to snow-albedo feedback. *Geophys. Res. Lett.* **45** 13–050.

BRACEGIRDLE, T. J. and STEPHENSON, D. B. (2012). Higher precision estimates of regional polar warming by ensemble regression of climate model projections. *Clim. Dyn.* **39** 2805–2821.

CHANDLER, R. E. (2013). Exploiting strength, discounting weakness: Combining information from multiple climate simulators. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20120388, 19. MR3046246 https://doi.org/10.1098/rsta.2012.0388

Key words and phrases. Multimodel ensembles, climate model integration, Gaussian process regression, deep learning.

EYRING, V., BONY, S., MEEHL, G. A., SENIOR, C. A., STEVENS, B., STOUFFER, R. J. and TAYLOR, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9** 1937–1958.

FLATO, G., MAROTZKE, J., ABIODUN, B., BRACONNOT, P., CHOU, S. C., COLLINS, W., COX, P., DRIOUECH, F., EMORI, S. et al. (2014). Evaluation of climate models. In *Climate Change* 2013: *The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* 741–866. Cambridge Univ. Press, Cambridge.

FRICKO, O., HAVLIK, P., ROGELJ, J., KLIMONT, Z., GUSTI, M., JOHNSON, N., KOLP, P., STRUBEGGER, M., VALIN, H. et al. (2017). The marker quantification of the shared socioeconomic pathway 2: A middle-of-the-road scenario for the 21st century. *Glob. Environ. Change* **42** 251–267.

GARRIGA-ALONSO, A., RASMUSSEN, C. E. and AITCHISON, L. (2018). Deep convolutional networks as shallow gaussian processes. Preprint. Available at arXiv:1808.05587.

GHAFARIANZADEH, M. and MONTELEONI, C. (2013). Climate prediction via matrix completion. In *AAAI* (*Late-Breaking Developments*).

GIORGI, F. and MEARNS, L. O. (2002). Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. *J. Climate* **15** 1141–1158.

GIORGI, F. and MEARNS, L. O. (2003). Probability of regional climate change based on the reliability ensemble averaging (REA) method. *Geophys. Res. Lett.* **30**.

GIORGI, F., RAFFAELE, F. and COPPOLA, E. (2019). The response of precipitation characteristics to global warming from climate projections. *Earth Syst. Dyn.* **10** 73–89.

GLECKLER, P. J., TAYLOR, K. E. and DOUTRIAUX, C. (2008). Performance metrics for climate models. *J. Geophys. Res., Atmos.* **113**.

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

GNEITING, T. and 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

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

GREENE, A. M., GODDARD, L. and LALL, U. (2006). Probabilistic multimodel regional temperature change projections. *J. Climate* **19** 4326–4343.

HARRIS, T., LI, B. and SRIVER, R. (2023). Supplement to "Multimodel ensemble analysis with neural network Gaussian processes." https://doi.org/10.1214/23-AOAS1768SUPPA, https://doi.org/10.1214/23-AOAS1768SUPPB

HAUGEN, M. A., STEIN, M. L., MOYER, E. J. and SRIVER, R. L. (2018). Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. *J. Climate* **31** 8573–8588.

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

KALNAY, E., KANAMITSU, M., KISTLER, R., COLLINS, W., DEAVEN, D., GANDIN, L., IREDELL, M., SAHA, S., WHITE, G. et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77** 437–472.

KATZFUSS, M. (2017). A multi-resolution approximation for massive spatial datasets. *J. Amer. Statist. Assoc.* **112** 201–214. MR3646566 https://doi.org/10.1080/01621459.2015.1123632

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KNUTTI, R., FURRER, R., TEBALDI, C., CERMAK, J. and MEEHL, G. A. (2010). Challenges in combining projections from multiple climate models. *J. Climate* **23** 2739–2758.

KNUTTI, R., SEDLÁČEK, J., SANDERSON, B. M., LORENZ, R., FISCHER, E. M. and EYRING, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophys. Res. Lett.* **44** 1909–1918.

KRISTIADI, A., HEIN, M. and HENNIG, P. (2020). Being Bayesian, even just a bit, fixes overconfidence in relu networks. In *International Conference on Machine Learning* 5436–5446. PMLR.

LAMBERT, S. J. and BOER, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. *Clim. Dyn.* **17** 83–106.

LEE, J., BAHRI, Y., NOVAK, R., SCHOENHOLZ, S. S., PENNINGTON, J. and SOHL-DICKSTEIN, J. (2017). Deep neural networks as gaussian processes. Preprint. Available at arXiv:1711.00165.

LENSSEN, N. J., GODDARD, L. and MASON, S. (2020). Seasonal forecast skill of ENSO teleconnection maps. *Weather Forecast.* **35** 2387–2406.

MacKay, D. C. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Comput*. **4** 448–472.

Mearns, L., McGinnis, S., Korytina, D., Arritt, R., Biner, S., Bukovsky, M., Chang, H., Christensen, O., Herzmann, D. et al. (2017). The NA-CORDEX dataset, version 1.0. NCAR Climate Data Gateway. Boulder (CO): The North American CORDEX Program 10. D6SJ1JCH.

Neal, R. M. (2012). *Bayesian Learning for Neural Networks* **118**. Springer, Berlin.

North, G. R., Bell, T. L., Cahalan, R. F. and Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review* **110** 699–706.

O'Neill, B. C., Tebaldi, C., Vuuren, D. P. v., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F. et al. (2016). The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development* **9** 3461–3482.

Polson, N. G. and Sokolov, V. (2017). Deep learning: A Bayesian perspective. *Bayesian Anal*. **12** 1275–1304. MR3724986 https://doi.org/10.1214/17-BA1082

Räisänen, J., Ruokolainen, L. and Ylhäisi, J. (2010). Weighting of model results for improving best estimates of climate change. *Climate Dynamics* **35** 407–422.

Ramachandran, P., Zoph, B. and Le, Q. V. (2017). Searching for activation functions. Preprint. Available at arXiv:1710.05941.

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

Rougier, J., Goldstein, M. and House, L. (2013). Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Statist. Assoc*. **108** 852–863. MR3174668 https://doi.org/10.1080/01621459.2013.802963

Sansom, P. G., Stephenson, D. B. and Bracegirdle, T. J. (2021). On constraining projections of future climate using observations and simulations from multiple climate models. *J. Amer. Statist. Assoc*. **116** 546–557. MR4270002 https://doi.org/10.1080/01621459.2020.1851696

Shand, L. and Li, B. (2017). Modeling nonstationarity in space and time. *Biometrics* **73** 759–768. MR3713110 https://doi.org/10.1111/biom.12656

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

Smith, R. L., Tebaldi, C., Nychka, D. and Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *J. Amer. Statist. Assoc*. **104** 97–116. MR2663036 https://doi.org/10.1198/jasa.2009.0007

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci*. **365** 2053–2075. MR2317897 https://doi.org/10.1098/rsta.2007.2076

Tebaldi, C., Mearns, L. O., Nychka, D. and Smith, R. L. (2004). Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters* **31**.

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Climate Research* **47** 123–138.

van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res*. **12** 2095–2119. MR2819028

Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters* **9** 81–84.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3426–3449
https://doi.org/10.1214/23-AOAS1769

BINNED MULTINOMIAL LOGISTIC REGRESSION FOR INTEGRATIVE CELL-TYPE ANNOTATION

BY KESHAV MOTWANI[1,a], RHONDA BACHER[2,b] AND AARON J. MOLSTAD[3,c]

[1]*Department of Biostatistics, University of Washington,* [a]*kmotwani@uw.edu*
[2]*Department of Biostatistics, University of Florida,* [b]*rbacher@ufl.edu*
[3]*Department of Statistics, University of Florida,* [c]*amolstad@ufl.edu*

Categorizing individual cells into one of many known cell-type categories, also known as cell-type annotation, is a critical step in the analysis of single-cell genomics data. The current process of annotation is time intensive and subjective, which has led to different studies describing cell types with labels of varying degrees of resolution. While supervised learning approaches have provided automated solutions to annotation, there remains a significant challenge in fitting a unified model for multiple datasets with inconsistent labels. In this article we propose a new multinomial logistic regression estimator which can be used to model cell-type probabilities by integrating multiple datasets with labels of varying resolution. To compute our estimator, we solve a nonconvex optimization problem using a blockwise proximal gradient descent algorithm. We show through simulation studies that our approach estimates cell-type probabilities more accurately than competitors in a wide variety of scenarios. We apply our method to 10 single-cell RNA-seq datasets and demonstrate its utility in predicting fine resolution cell-type labels on unlabeled data as well as refining cell-type labels on data with existing coarse resolution annotations. Finally, we demonstrate that our method can lead to novel scientific insights in the context of a differential expression analysis comparing peripheral blood gene expression before and after treatment with interferon-β. An R package implementing the method is available in the Supplementary Material and at https://github.com/keshav-motwani/IBMR, and the collection of datasets we analyze is available at https://github.com/keshav-motwani/AnnotatedPBMC.

REFERENCES

10X GENOMICS (2018). 10k PBMCs from a healthy donor—gene expression and cell surface protein. Available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3.

10X GENOMICS (2019). 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry). Available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_protein_v3.

ABDELAAL, T., MICHIELSEN, L., CATS, D., HOOGDUIN, D., MEI, H., REINDERS, M. J. and MAHFOUZ, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20** 194.

AITKEN, S., MAGI, S., ALHENDI, A. M., ITOH, M., KAWAJI, H., LASSMANN, T., DAUB, C. O., ARNER, E., CARNINCI, P. et al. (2015). Transcriptional dynamics reveal critical roles for non-coding RNAs in the immediate-early response. *PLoS Comput. Biol.* **11** e1004217.

AMEZQUITA, R. A., LUN, A. T., BECHT, E., CAREY, V. J., CARPP, L. N., GEISTLINGER, L., MARINI, F., RUE-ALBRECHT, K., RISSO, D. et al. (2020). Orchestrating single-cell analysis with bioconductor. *Nat. Methods* **17** 137–145.

ARAN, D., LOONEY, A. P., LIU, L., WU, E., FONG, V., HSU, A., CHAK, S., NAIKAWADI, R. P., WOLTERS, P. J. et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20** 163–172.

Key words and phrases. Integrative analysis, multinomial logistic regression, group lasso, nonconvex optimization, single-cell genomics, cell-type annotation.

BARRY, C., SCHMITZ, M. T., ARGUS, C., BOLIN, J. M., PROBASCO, M. D., LENG, N., DUFFIN, B. M., STEILL, J., SWANSON, S. et al. (2019). Automated minute scale RNA-seq of pluripotent stem cell differentiation reveals early divergence of human and mouse gene expression kinetics. *PLoS Comput. Biol.* **15** e1007543.

BERARD, M. and TOUGH, D. F. (2002). Qualitative differences between naive and memory T cells. *Immunology* **106** 127.

CONDE, C. D., GOMES, T., JARVIS, L. B., XU, C., HOWLETT, S., RAINBOW, D., SUCHANEK, O., KING, H., MAMANOVA, L. et al. (2021). Cross-tissue immune cell analysis reveals tissue-specific adaptations and clonal architecture across the human body. *bioRxiv*.

CROWELL, H. L., SONESON, C., GERMAIN, P.-L., CALINI, D., COLLIN, L., RAPOSO, C., MALHOTRA, D. and ROBINSON, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11** 1–12.

DENG, Y., HUANG, Z., ZHOU, C., WANG, J., YOU, Y., SONG, Z., XIANG, M., ZHONG, B. and HAO, F. (2006). Gene profiling involved in immature CD4+ T lymphocyte responsible for systemic lupus erythematosus. *Mol. Immunol.* **43** 1497–1507.

DING, J., ADICONIS, X., SIMMONS, S. K., KOWALCZYK, M. S., HESSION, C. C., MARJANOVIC, N. D., HUGHES, T. K., WADSWORTH, M. H., BURKS, T. et al. (2019). Systematic comparative analysis of single cell RNA-sequencing methods. *BioRxiv* 632216.

DOZMOROV, I., DOMINGUEZ, N., SESTAK, A. L., ROBERTSON, J. M., HARLEY, J. B., JAMES, J. A. and GUTHRIDGE, J. M. (2013). Evidence of dynamically dysregulated gene expression pathways in hyperresponsive B cells from African American lupus patients. *PLoS ONE* **8** e71397.

HAGHVERDI, L., LUN, A. T., MORGAN, M. D. and MARIONI, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36** 421–427.

HAO, Y., HAO, S., ANDERSEN-NISSEN, E., MAUCK, W. M., ZHENG, S., BUTLER, A., LEE, M. J., WILK, A. J., DARBY, C. et al. (2020). Integrated analysis of multimodal single-cell data. *bioRxiv*.

HIE, B., BRYSON, B. and BERGER, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37** 685–691.

HORVATH, C. M. (2004). The Jak-STAT pathway stimulated by interferon α or interferon β. *Science's STKE* **260** tr10–tr10.

HUANG, Q., LIU, Y., DU, Y. and GARMIRE, L. X. (2021). Evaluation of cell type annotation R packages on single-cell RNA-seq data. *Genomics Proteomics Bioinform.* **19** 267–281.

HUANG, Y., ZHANG, Q., ZHANG, S., HUANG, J. and MA, S. (2017). Promoting similarity of sparsity structures in integrative analysis with penalization. *J. Amer. Statist. Assoc.* **112** 342–350. MR3646576 https://doi.org/10.1080/01621459.2016.1139497

KANG, H. M., SUBRAMANIAM, M., TARG, S., NGUYEN, M., MALISKOVA, L., MCCARTHY, E., WAN, E., WONG, S., BYRNES, L. et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36** 89–94.

KORSUNSKY, I., MILLARD, N., FAN, J., SLOWIKOWSKI, K., ZHANG, F., WEI, K., BAGLAENKO, Y., BRENNER, M., LOH, P.-R. et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* **16** 1289–1296.

KOTLIAROV, Y., SPARKS, R., MARTINS, A. J., MULÈ, M. P., LU, Y., GOSWAMI, M., KARDAVA, L., BANCHEREAU, R., PASCUAL, V. et al. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* **26** 618–629.

LÄHNEMANN, D., KÖSTER, J., SZCZUREK, E., MCCARTHY, D. J., HICKS, S. C., ROBINSON, M. D., VALLEJOS, C. A., CAMPBELL, K. R., BEERENWINKEL, N. et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* **21** 1–35.

LANGE, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia, PA. MR3522165 https://doi.org/10.1137/1.9781611974409.ch1

LIU, C., MARTINS, A. J., LAU, W. W., RACHMANINOFF, N., CHEN, J., IMBERTI, L., MOSTAGHIMI, D., FINK, D. L., BURBELO, P. D. et al. (2021). Time-resolved systems immunology reveals a late juncture linked to fatal Covid-19. *Cell* **184** 1836–1857.

LUECKEN, M. D., BÜTTNER, M., CHAICHOOMPU, K., DANESE, A., INTERLANDI, M., MÜLLER, M. F., STROBL, D. C., ZAPPIA, L., DUGAS, M. et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19** 41–50.

MA, W., SU, K. and WU, H. (2021). Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: Classifier, feature selection, and reference construction. *Genome Biol.* **22** 1–23.

MOLSTAD, A. J. and PATRA, R. K. (2022). Dimension reduction for integrative survival analysis. *Biometrics*. https://doi.org/10.1111/biom.13736

MOLSTAD, A. J. and ROTHMAN, A. J. (2023). A likelihood-based approach for multivariate categorical response regression in high dimensions. *J. Amer. Statist. Assoc.* **118** 1402–1414. MR4595503 https://doi.org/10.1080/01621459.2021.1999819

MOTWANI, K., BACHER, R. and MOLSTAD, A. J. (2023). Supplement to "Binned multinomial logistic regression for integrative cell-type annotation." https://doi.org/10.1214/23-AOAS1769SUPPA, https://doi.org/10.1214/23-AOAS1769SUPPB

NEEB, A., WALLBAUM, S., NOVAC, N., DUKOVIC-SCHULZE, S., SCHOLL, I., SCHREIBER, C., SCHLAG, P., MOLL, J., STEIN, U. et al. (2012). The immediate early gene IER2 promotes tumor cell motility and metastasis, and predicts poor survival of colorectal cancer patients. *Oncogene* **31** 3796–3806.

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839 https://doi.org/10.1214/09-AOS776

OSHLACK, A., ROBINSON, M. D. and YOUNG, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biol.* **11** 1–10.

PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

PASQUINI, G., ROJO ARIAS, J. E., SCHÄFER, P. and BUSSKAMP, V. (2021). Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19** 961–969.

POLSON, N. G., SCOTT, J. G. and WILLARD, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statist. Sci.* **30** 559–581. MR3432841 https://doi.org/10.1214/15-STS530

SCHAUM, N., KARKANIAS, J., NEFF, N. F., MAY, A. P., QUAKE, S. R., WYSS-CORAY, T., DARMANIS, S., BATSON, J., BOTVINNIK, O. et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature* **562** 367.

SHASHA, C., TIAN, Y., MAIR, F., MILLER, H. E. and GOTTARDO, R. (2021). Superscan: Supervised single-cell annotation. *bioRxiv*.

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

STEPHENSON, E., REYNOLDS, G., BOTTING, R. A., CALERO-NIETO, F. J., MORGAN, M. D., TUONG, Z. K., BACH, K., SUNGNAK, W., WORLOCK, K. B. et al. (2021). Single-cell multi-omics analysis of the immune response in Covid-19. *Nat. Med.* **27** 904–916.

SU, Y., CHEN, D., YUAN, D., LAUSTED, C., CHOI, J., DAI, C. L., VOILLET, V., DUVVURI, V. R., SCHERLER, K. et al. (2020). Multi-omics resolves a sharp disease-state shift between mild and moderate Covid-19. *Cell* **183** 1479–1495.

VENTZ, S., MAZUMDER, R. and TRIPPA, L. (2022). Integration of survival data from multiple studies. *Biometrics* **78** 1365–1376. MR4534363 https://doi.org/10.1111/biom.13517

WILK, A. J., RUSTAGI, A., ZHAO, N. Q., ROQUE, J., MARTÍNEZ-COLÓN, G. J., MCKECHNIE, J. L., IVISON, G. T., RANGANATH, T., VERGARA, R. et al. (2020). A single-cell atlas of the peripheral immune response in patients with severe Covid-19. *Nat. Med.* **26** 1070–1076.

WOLF, F. A., ANGERER, P. and THEIS, F. J. (2018). Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19** 1–5.

XIE, Z., BAILEY, A., KULESHOV, M. V., CLARKE, D. J., EVANGELISTA, J. E., JENKINS, S. L., LACHMANN, A., WOJCIECHOWICZ, M. L., KROPIWNICKI, E. et al. (2021). Gene set knowledge discovery with enrichr. *Curr. Protoc.* **1** e90.

XU, Y. and YIN, W. (2017). A globally convergent algorithm for nonconvex optimization based on block coordinate update. *J. Sci. Comput.* **72** 700–734. MR3673692 https://doi.org/10.1007/s10915-017-0376-0

YARILINA, A. and IVASHKIV, L. B. (2010). Type I interferon: A new player in TNF signaling. *TNF Pathophysiol.* **11** 94–104.

YOUNG, M. D. and BEHJATI, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9** giaa151.

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

ZHAO, Q., SHI, X., HUANG, J., LIU, J., LI, Y. and MA, S. (2015). Integrative analysis of '-omics' data using penalty functions. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 99–108. MR3348725 https://doi.org/10.1002/wics.1322

ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R., ZIRALDO, S. B., WHEELER, T. D., MCDERMOTT, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 1–12.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3450–3475
https://doi.org/10.1214/23-AOAS1771

COMPRESSED SPECTRAL SCREENING FOR LARGE-SCALE DIFFERENTIAL CORRELATION ANALYSIS WITH APPLICATION IN SELECTING GLIOBLASTOMA GENE MODULES

BY TIANXI LI[1,a], XIWEI TANG[1,b] AND AJAY CHATRATH[2,c]

[1]*Department of Statistics, University of Virginia,* [a]*tianxili@virginia.edu,* [b]*xt4yj@virginia.edu*
[2]*Department of Neurosurgery, Washington University School of Medicine in St. Louis,* [c]*achatrath@wustl.edu*

Differential coexpression analysis has been widely applied by scientists in understanding the biological mechanisms of diseases. However, the unknown differential patterns are often complicated; thus, models based on simplified parametric assumptions can be ineffective in identifying the differences. Meanwhile, the gene expression data involved in such analysis are in extremely high dimensions by nature, whose correlation matrices may not even be computable. Such a large scale seriously limits the application of most well-studied statistical methods. This paper introduces a simple yet powerful approach to the differential correlation analysis problem called compressed spectral screening. By leveraging spectral structures and random sampling techniques, our approach could achieve a highly accurate screening of features with complicated differential patterns while maintaining the scalability to analyze correlation matrices of 10^4–10^5 variables within a few minutes on a standard personal computer. We have applied this screening approach in comparing a TCGA data set about Glioblastoma with normal subjects. Our analysis successfully identifies multiple functional modules of genes that exhibit different coexpression patterns. The findings reveal new insights about Glioblastoma's evolving mechanism. The validity of our approach is also justified by a theoretical analysis, showing that the compressed spectral analysis can achieve variable screening consistency.

REFERENCES

ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

ANANDKUMAR, A., FOSTER, D. P., HSU, D., KAKADE, S. M. and LIU, Y.-K. (2015). A spectral algorithm for latent Dirichlet allocation. *Algorithmica* **72** 193–214. MR3332930 https://doi.org/10.1007/s00453-014-9909-1

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Nat. Prec.* 1–1.

ARCHER, S. K., SHIROKIKH, N. E., HALLWIRTH, C. V., BEILHARZ, T. H. and PREISS, T. (2015). Probing the closed-loop model of mRNA translation in living cells. *RNA Biol.* **12** 248–254. https://doi.org/10.1080/15476286.2015.1017242

ARDLIE, K. G., DELUCA, D. S., SEGRÈ, A. V., SULLIVAN, T. J., YOUNG, T. R. et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** 648–660. https://doi.org/10.1126/science.1262110

BACOLOD, M. D. and BARANY, F. (2021). MGMT epigenetics: The influence of gene body methylation and other insights derived from integrated methylomic, transcriptomic, and chromatin analyses in various cancer types. *Curr Cancer Drug Targets* **21** 360–374. https://doi.org/10.2174/1568009621666210203111620

BALLOUZ, S., VERLEYEN, W. and GILLIS, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers. *Bioinformatics* **31** 2123–2130. https://doi.org/10.1093/bioinformatics/btv118

BARABÁSI, A.-L. (2009). Scale-free networks: A decade and beyond. *Science* **325** 412–413. MR2548299 https://doi.org/10.1126/science.1173299

Key words and phrases. Differential correlation analysis, spectral methods, high-dimensional correlation matrices, gene coexpression.

BARABÁSI, A.-L. and BONABEAU, E. (2003). Scale-free networks. *Sci. Amer.* **288** 60–69. https://doi.org/10.1038/scientificamerican0503-60

BHUVA, D. D., CURSONS, J., SMYTH, G. K. and DAVIS, M. J. (2019). Differential coexpression-based detection of conditional relationships in transcriptional data: Comparative analysis and application to breast cancer. *Genome Biol.* **20** 1–21.

BRAT, D. J., VERHAAK, R. G., SALAMA, S. R., COOPER, L. et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372** 2481–2498. https://doi.org/10.1056/NEJMoa1402121

BRINGMANN, K. and FRIEDRICH, T. (2013). Exact and efficient generation of geometric random variates and random graphs. In *Automata, Languages, and Programming. Part I. Lecture Notes in Computer Science* **7965** 267–278. Springer, Heidelberg. MR3109077 https://doi.org/10.1007/978-3-642-39206-1_23

BULLARD, J. H., PURDOM, E., HANSEN, K. D. and DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinform.* **11** 1–13.

BUTUCEA, C., INGSTER, Y. I. and SUSLINA, I. A. (2015). Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM Probab. Stat.* **19** 115–134. MR3374872 https://doi.org/10.1051/ps/2014017

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

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618 https://doi.org/10.1080/01621459.2012.758041

CAI, T. T., LIANG, T. and RAKHLIN, A. (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Ann. Statist.* **45** 1403–1430. MR3670183 https://doi.org/10.1214/16-AOS1488

CAI, T. T. and ZHANG, A. (2016). Inference for high-dimensional differential correlation matrices. *J. Multivariate Anal.* **143** 107–126. MR3431422 https://doi.org/10.1016/j.jmva.2015.08.019

CAI, Z., LI, R. and ZHANG, Y. (2022). A distribution free conditional independence test with applications to causal discovery. *J. Mach. Learn. Res.* **23** Paper No. [85], 41. MR4576670

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5

CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Ann. Statist.* **47** 2405–2439. MR3988761 https://doi.org/10.1214/18-AOS1752

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

CHANG, J., ZHOU, W., ZHOU, W.-X. and WANG, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics* **73** 31–41. MR3632349 https://doi.org/10.1111/biom.12552

CHATRATH, A., PRZANOWSKA, R., KIRAN, S., SU, Z., SAHA, S., WILSON, B., TSUNEMATSU, T., AHN, J.-H., LEE, K. Y. et al. (2020). The pan-cancer landscape of prognostic germline variants in 10,582 patients. *Gen. Med.* **12** 1–18.

CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

CHEN, Y., CHI, Y., FAN, J., MA, C. and YAN, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM J. Optim.* **30** 3098–3121. MR4167625 https://doi.org/10.1137/19M1290000

CHEN, Y. and XU, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.* **17** Paper No. 27, 57. MR3491121

CHI, E. C. and LI, T. (2019). Matrix completion from a computational statistics perspective. *Wiley Interdiscip. Rev.: Comput. Stat.* **11** e1469, 25. MR3999530 https://doi.org/10.1002/wics.1469

CHIQUET, J., GRANDVALET, Y. and AMBROISE, C. (2011). Inferring multiple graphical structures. *Stat. Comput.* **21** 537–553. MR2826691 https://doi.org/10.1007/s11222-010-9191-2

CIECHOMSKA, I. A., JAYAPRAKASH, C., MALESZEWSKA, M. and KAMINSKA, B. (2020). Histone modifying enzymes and chromatin modifiers in glioma pathobiology and therapy responses. *Adv. Exp. Med. Biol.* **1202** 259–279. https://doi.org/10.1007/978-3-030-30651-9_13

COSTA-SILVA, J., DOMINGUES, D. and LOPES, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **12** e0190152.

DADANEH, S. Z., QIAN, X. and ZHOU, M. (2018). BNP-Seq: Bayesian nonparametric differential expression analysis of sequencing count data. *J. Amer. Statist. Assoc.* **113** 81–94. MR3803441 https://doi.org/10.1080/01621459.2017.1328358

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871 https://doi.org/10.1111/rssb.12033

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

EL-KHAYAT, S. M. and ARAFAT, W. O. (2021). Therapeutic strategies of recurrent glioblastoma and its molecular pathways 'Lock up the beast'. *Ecancermedicalscience* **15** 1176. https://doi.org/10.3332/ecancer.2021.1176

ENGELHARDT, B. E. and STEPHENS, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* **6** e1001117. https://doi.org/10.1371/journal.pgen.1001117

ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031

FAN, J., WANG, K., ZHONG, Y. and ZHU, Z. (2021). Robust high-dimensional factor models with applications to statistical machine learning. *Statist. Sci.* **36** 303–327. MR4255196 https://doi.org/10.1214/20-sts785

FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** Paper No. 207, 42. MR3827095

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

GAMBARDELLA, G., MORETTI, M. N., DE CEGLI, R., CARDONE, L., PERON, A. and DI BERNARDO, D. (2013). Differential network analysis for the identification of condition-specific pathway activity and regulation. *Bioinformatics* **29** 1776–1785. https://doi.org/10.1093/bioinformatics/btt290

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206 https://doi.org/10.1093/biomet/asq060

HANSEN, K. D., IRIZARRY, R. A. and WU, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13** 204–216.

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

HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.* **78** 1460–1480. MR2926144 https://doi.org/10.1016/j.jcss.2011.12.025

HUDSON, N. J., REVERTER, A. and DALRYMPLE, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.* **5** e1000382. https://doi.org/10.1371/journal.pcbi.1000382

JIA, L., SONG, H., FAN, W., SONG, Y., WANG, G., LI, X., HE, Y. and YAO, A. (2020). The association between high mobility group box 1 chromatin protein and mitotic chromosomes in glioma cells. *Oncol. Lett.* **19** 745–752.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

KAI, Z., DINGYANG, L. and ZHUANYI, Y. (2021). Prognostic role of BRAF mutation in low-grade gliomas: Meta-analysis. *World Neurosurg.* **147** 42–46. https://doi.org/10.1016/j.wneu.2020.12.029

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

KRUPP, M., MARQUARDT, J. U., SAHIN, U., GALLE, P. R., CASTLE, J. and TEUFEL, A. (2012). RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* **28** 1184–1185. https://doi.org/10.1093/bioinformatics/bts084

LE, C. M. and LI, T. (2022). Linear regression and its inference on noisy network-linked data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1851–1885. MR4515560

LEI, J. (2020). Cross-validation with confidence. *J. Amer. Statist. Assoc.* **115** 1978–1997. MR4189771 https://doi.org/10.1080/01621459.2019.1672556

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, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27** 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938 https://doi.org/10.1214/12-AOS993

LI, J. and TSENG, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* **5** 994–1019. MR2840184 https://doi.org/10.1214/10-AOAS393

LI, J.-R., SUN, C.-H., LI, W., CHAO, R.-F., HUANG, C.-C., ZHOU, X. J. and LIU, C.-C. (2016). Cancer RNA-Seq Nexus: A database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Res*. **44** D944–D951.

LI, T., LEI, L., BHATTACHARYYA, S., VAN DEN BERGE, K., SARKAR, P., BICKEL, P. J. and LEVINA, E. (2022). Hierarchical community detection by recursive partitioning. *J. Amer. Statist. Assoc*. **117** 951–968. MR4436325 https://doi.org/10.1080/01621459.2020.1833888

LI, T., LEVINA, E. and ZHU, J. (2020a). Community models for networks observed through edge nominations. Preprint. Available at arXiv:2008.03652.

LI, T., LEVINA, E. and ZHU, J. (2020b). Network cross-validation by edge sampling. *Biometrika* **107** 257–276. MR4108931 https://doi.org/10.1093/biomet/asaa006

LI, T., TANG, X. and CHATRATH, A. (2023). Supplement to "Compressed spectral screening for large-scale differential correlation analysis with application in selecting Glioblastoma gene modules." https://doi.org/10.1214/23-AOAS1771SUPPA, https://doi.org/10.1214/23-AOAS1771SUPPB

LIN, L., DRTON, M. and SHOJAIE, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat*. **10** 806–854. MR3486418 https://doi.org/10.1214/16-EJS1126

LIU, Y. and ARIAS-CASTRO, E. (2019). A multiscale scan statistic for adaptive submatrix localization. In *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 44–53.

LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet*. **45** 580–585.

LOUIS, D. N., PERRY, A., REIFENBERGER, G., VON DEIMLING, A., FIGARELLA-BRANGER, D., CAVENEE, W. K., OHGAKI, H., WIESTLER, O. D., KLEIHUES, P. et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: A summary. *Acta Neuropathol*. **131** 803–820. https://doi.org/10.1007/s00401-016-1545-1

MARKO, N. F. and WEIL, R. J. (2012). Non-Gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PLoS ONE* **7** e46935. https://doi.org/10.1371/journal.pone.0046935

MASAYESVA, B. G., HA, P., GARRETT-MAYER, E., PILKINGTON, T., MAO, R., PEVSNER, J., SPEED, T., BENOIT, N., MOON, C.-S. et al. (2004). Gene expression alterations over large chromosomal regions in cancers include multiple genes unrelated to malignant progression. *Proc. Natl. Acad. Sci. USA* **101** 8715–8720.

MAYRINK, V. D. and LUCAS, J. E. (2013). Sparse latent factor models with interactions: Analysis of gene expression data. *Ann. Appl. Stat*. **7** 799–822. MR3112918 https://doi.org/10.1214/12-AOAS607

MCKENZIE, A. T., KATSYV, I., SONG, W.-M., WANG, M. and ZHANG, B. (2016). DGCA: A comprehensive R package for differential gene correlation analysis. *BMC Syst. Biol*. **10** 1–25.

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

MI, H., MURUGANUJAN, A., EBERT, D., HUANG, X. and THOMAS, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. **47** D419–D426.

MIAO, R. and LI, T. (2023). Informative core identification in complex networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol*.. qkac009. https://doi.org/10.1093/jrsssb/qkac009

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.

RISSO, D., NGAI, J., SPEED, T. P. and DUDOIT, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol*. **32** 896–902.

RISSO, D., SCHWARTZ, K., SHERLOCK, G. and DUDOIT, S. (2011). GC-content normalization for RNA-seq data. *BMC Bioinform*. **12** 1–17.

ROBINSON, M. D. and OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11** 1–9.

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

RUNCIE, D. E. and MUKHERJEE, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics* **194** 753–767.

SAAD, Y. (2011). *Numerical Methods for Large Eigenvalue Problems*. *Classics in Applied Mathematics* **66**. SIAM, Philadelphia, PA. MR3396212 https://doi.org/10.1137/1.9781611970739.ch1

SAEGUSA, T. and SHOJAIE, A. (2016). Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Stat*. **10** 1341–1392. MR3507368 https://doi.org/10.1214/16-EJS1137

SANCHEZ-CASTILLO, M., TIENDA-LUNA, I., BLANCO, D., CARRION-PEREZ, M. and HUANG, Y. (2013). Bayesian sparse factor model for transcriptional regulatory networks inference. In 21*st European Signal Processing Conference* (*EUSIPCO* 2013) 1–4. IEEE, New York.

SARMAH, T. and BHATTACHARYYA, D. K. (2021). A study of tools for differential coexpression analysis for RNA-seq data. *Inform. Med. Unlocked* 100740.

SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* **51** 6535–6542. MR2408613 https://doi.org/10.1016/j.csda.2007.03.004

SHABALIN, A. A., WEIGMAN, V. J., PEROU, C. M. and NOBEL, A. B. (2009). Finding large average submatrices in high dimensional data. *Ann. Appl. Stat.* **3** 985–1012. MR2750383 https://doi.org/10.1214/09-AOAS239

SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

SHI, T., GUO, D., XU, H., SU, G., CHEN, J., ZHAO, Z., SHI, J., WEDEMEYER, M., ATTENELLO, F. et al. (2020). HOTAIRM1, an enhancer lncRNA, promotes glioma proliferation by regulating long-range chromatin interactions within HOXA cluster genes. *Mol. Biol. Rep.* 1–11.

SHOJAIE, A. (2021). Differential network analysis: A statistical perspective. *Wiley Interdiscip. Rev.: Comput. Stat.* **13** Paper No. e1508, 16. MR4218944 https://doi.org/10.1002/wics.1508

SISKA, C., BOWLER, R. and KECHRIS, K. (2016). The discordant method: A novel approach for differential correlation. *Bioinformatics* **32** 690–696.

SISKA, C. and KECHRIS, K. (2020). discordant: The discordant method: A novel approach for differential correlation. R package version 1.14.0.

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **14** 1–18.

SUN, W., LIU, Y., CROWLEY, J. J. et al. (2015). IsoDOT detects differential RNA-isoform expression/usage with respect to a categorical or continuous covariate with high sensitivity and specificity. *J. Amer. Statist. Assoc.* **110** 975–986. MR3420677 https://doi.org/10.1080/01621459.2015.1040880

TARAZONA, S., GARCÍA-ALCALDE, F., DOPAZO, J., FERRER, A. and CONESA, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21** 2213–2223.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7** 562–578.

VAN DER WIJST, M. G., BRUGGE, H., DE VRIES, D. H., DEELEN, P., SWERTZ, M. A. and FRANKE, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and coexpression QTLs. *Nat. Genet.* **50** 493–497.

WAN, Q., DINGERDISSEN, H., FAN, Y., GULZAR, N., PAN, Y., WU, T.-J., YAN, C., ZHANG, H. and MAZUMDER, R. (2015). BioXpress: An integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database* **2015**.

WANG, H. and HE, X. (2007). Detecting differential expressions in GeneChip microarray studies: A quantile approach. *J. Amer. Statist. Assoc.* **102** 104–112. MR2293303 https://doi.org/10.1198/016214506000001220

WANG, Q., ARMENIA, J., ZHANG, C., PENSON, A. V., REZNIK, E., ZHANG, L., MINET, T., OCHOA, A., GROSS, B. E. et al. (2018). Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* **5** 1–8.

XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102** 247–266. MR3371002 https://doi.org/10.1093/biomet/asu074

XUE, W., KITZING, T., ROESSLER, S., ZUBER, J., KRASNITZ, A., SCHULTZ, N., REVILL, K., WEISS-MUELLER, S., RAPPAPORT, A. R. et al. (2012). A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl. Acad. Sci. USA* **109** 8212–8217.

YAN, W., ZHANG, W., YOU, G., ZHANG, J., HAN, L., BAO, Z., WANG, Y., LIU, Y., JIANG, C. et al. (2012). Molecular classification of gliomas based on whole genome gene expression: A systematic report of 225 samples from the Chinese Glioma Cooperative Group. *J. Neuro-Oncol.* **14** 1432–1440.

YANG, J., LIU, Y., LIU, Y. and SUN, W. (2021). Model free estimation of graphical model using gene expression data. *Ann. Appl. Stat.* **15** 194–207. MR4255264 https://doi.org/10.1214/20-aoas1380

YU, M., GUPTA, V. and KOLAR, M. (2020). Simultaneous inference for pairwise graphical models with generalized score matching. *J. Mach. Learn. Res.* **21** Paper No. 91, 51. MR4119159

YUAN, H., XI, R., CHEN, C. and DENG, M. (2017). Differential network analysis via lasso penalized D-trace loss. *Biometrika* **104** 755–770. MR3737302 https://doi.org/10.1093/biomet/asx049

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

ZHANG, B. and MCKENZIE, A. (2019). DGCA: Differential Gene Correlation Analysis. R package version 1.0.2.

ZHANG, Y., LIN, Z., LIN, X., ZHANG, X., ZHAO, Q. and SUN, Y. (2021). A gene module identification algorithm and its applications to identify gene modules and key genes of hepatocellular carcinoma. *Sci. Rep.* **11** 5517.

ZHANG, Z. H., JHAVERI, D. J., MARSHALL, V. M., BAUER, D. C., EDSON, J., NARAYANAN, R. K., ROBINSON, G. J., LUNDBERG, A. E., BARTLETT, P. F. et al. (2014). A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE* **9** e103207.

ZHAO, S. and SHOJAIE, A. (2022). Network differential connectivity analysis. *Ann. Appl. Stat.* **16** 2166–2182. MR4489204 https://doi.org/10.1214/21-aoas1581

ZHAO, S. D., CAI, T. T., CAPPOLA, T. P., MARGULIES, K. B. and LI, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *J. Amer. Statist. Assoc.* **112** 1032–1046. MR3735358 https://doi.org/10.1080/01621459.2016.1270825

ZHAO, S. D., CAI, T. T. and LI, H. (2014). Direct estimation of differential networks. *Biometrika* **101** 253–268. MR3215346 https://doi.org/10.1093/biomet/asu009

ZHAO, Z., ZHANG, K.-N., WANG, Q., LI, G., ZENG, F., ZHANG, Y., WU, F., CHAI, R., WANG, Z. et al. (2021). Chinese Glioma Genome Atlas (CGGA): A comprehensive resource with functional genomic data from Chinese glioma patients. *Genomics Proteomics Bioinform.* **19** 1–12.

ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2017). Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. *Ann. Appl. Stat.* **11** 1810–1831. MR3709579 https://doi.org/10.1214/17-AOAS1062

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3476–3499
https://doi.org/10.1214/23-AOAS1772

A STATISTICAL APPROACH TO ESTIMATING ADSORPTION-ISOTHERM PARAMETERS IN GRADIENT-ELUTION PREPARATIVE LIQUID CHROMATOGRAPHY

BY JIAJI SU[1,a], ZHIGANG YAO[1,b], CHENG LI[1,c] AND YE ZHANG[2,d]

[1]*Department of Statistics and Data Science, National University of Singapore,* [a]*su_jiaji@u.nus.edu,* [b]*zhigang.yao@nus.edu.sg,* [c]*stalic@nus.edu.sg*

[2]*Faculty of Computational Mathematics and Cybernetics, Shenzhen MSU-BIT University,* [d]*ye.zhang@smbu.edu.cn*

Determining the adsorption isotherms is an issue of significant importance in preparative chromatography. A modern technique for estimating adsorption isotherms is to solve an inverse problem so that the simulated batch separation coincides with actual experimental results. However, due to the ill-posedness, the high nonlinearity, and the uncertainty quantification of the corresponding physical model, the existing deterministic inversion methods are usually inefficient in real-world applications. To overcome these difficulties and study the uncertainties of the adsorption-isotherm parameters, in this work, based on the Bayesian sampling framework, we propose a statistical approach for estimating the adsorption isotherms in various chromatography systems. Two modified Markov chain Monte Carlo algorithms are developed for a numerical realization of our statistical approach. Numerical experiments with both synthetic and real data are conducted and described to show the efficiency of the proposed new method.

REFERENCES

BELLONI, A. and CHERNOZHUKOV, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.* **37** 2011–2055. MR2533478 https://doi.org/10.1214/08-AOS634

BERRYMAN, J. G. and HOLLAND, C. J. (1978). Nonlinear diffusion problem arising in plasma physics. *Phys. Rev. Lett.* **40** 1720–1722. MR0495716 https://doi.org/10.1103/PhysRevLett.40.1720

CHAIKOVSKII, D. and ZHANG, Y. (2022). Convergence analysis for forward and inverse problems in singularly perturbed time-dependent reaction-advection-diffusion equations. *J. Comput. Phys.* **470** Paper No. 111609, 32. MR4486149 https://doi.org/10.1016/j.jcp.2022.111609

CHENG, X., LIN, G., ZHANG, Y., GONG, R. and GULLIKSSON, M. (2018). A modified coupled complex boundary method for an inverse chromatography problem. *J. Inverse Ill-Posed Probl.* **26** 33–49. MR3757483 https://doi.org/10.1515/jiip-2016-0057

CHKREBTII, O. A., CAMPBELL, D. A., CALDERHEAD, B. and GIROLAMI, M. A. (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Anal.* **11** 1239–1267. MR3577378 https://doi.org/10.1214/16-BA1017

COCKAYNE, J., OATES, C. J., SULLIVAN, T. J. and GIROLAMI, M. (2019). Bayesian probabilistic numerical methods. *SIAM Rev.* **61** 756–789. MR4027836 https://doi.org/10.1137/17M1139357

DOSE, E. V., JACOBSON, S. and GUIOCHON, G. (1991). Determination of isotherms from chromatographic peak shapes. *Anal. Chem.* **63** 833–839.

FELINGER, A., ZHOU, D. and GUIOCHON, G. (2003). Determination of the single component and competitive adsorption isotherms of the 1-indanol enantiomers by the inverse method. *J. Chromatogr. A* **1005** 35–49.

FORSSÉN, P., ARNELL, R. and FORNSTEDT, T. (2006). An improved algorithm for solving inverse problems in liquid chromatography. *Comput. Chem. Eng.* **30** 1381–1391.

GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

GUIOCHON, G. and LIN, B. (2003). *Modeling for Preparative Chromatography*. Academic Press, New York.

Key words and phrases. Liquid chromatography, adsorption isotherm, inverse problem, Bayesian sampling, Gaussian-mixture model.

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

HIDALGO, A., TELLO, L. and TORO, E. F. (2014). Numerical and analytical study of an atherosclerosis inflammatory disease model. *J. Math. Biol.* **68** 1785–1814. MR3201914 https://doi.org/10.1007/s00285-013-0688-0

HORVATH, C. (1988). *In High-Performance Liquid Chromatography: Advances and Perspectives* (*Volume* 5). Academic Press, New York.

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

JAVEED, S., QAMAR, S., SEIDEL-MORGENSTERN, A. and WARNECKE, G. (2011). Efficient and accurate numerical simulation of nonlinear chromatographic processes. *Comput. Chem. Eng.* **35** 2294–2305.

LIN, G., ZHANG, Y., CHENG, X., GULLIKSSON, M., FORSSÉN, P. and FORNSTEDT, T. (2018). A regularizing Kohn-Vogelius formulation for the model-free adsorption isotherm estimation problem in chromatography. *Appl. Anal.* **97** 13–40. MR3764747 https://doi.org/10.1080/00036811.2017.1284311

LISEC, O., HUGO, P. and SEIDEL-MORGENSTERN, A. (2001). Frontal analysis method to determine competitive adsorption isotherms. *J. Chromatogr. A* **908** 19–34. https://doi.org/10.1016/s0021-9673(00)00966-3

ROBERTS, G. O. and ROSENTHAL, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 255–268. MR1625691 https://doi.org/10.1111/1467-9868.00123

ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

RUTHVEN, D. M. (1984). *Principles of Adsorption and Adsorption Processes*. Wiley, New York.

SU, J., YAO, Z., LI, C. and ZHANG, Y. (2023). Supplement to "A statistical approach to estimating adsorption-isotherm parameters in gradient-elution preparative liquid chromatography." https://doi.org/10.1214/23-AOAS1772SUPPA, https://doi.org/10.1214/23-AOAS1772SUPPB

XUN, X., CAO, J., MALLICK, B., MAITY, A. and CARROLL, R. J. (2013). Parameter estimation of partial differential equation models. *J. Amer. Statist. Assoc.* **108** 1009–1020. MR3174680 https://doi.org/10.1080/01621459.2013.794730

ZHANG, Y., LIN, G., GULLIKSSON, M., FORSSÉN, P., FORNSTEDT, T. and CHENG, X. (2017). An adjoint method in inverse problems of chromatography. *Inverse Probl. Sci. Eng.* **25** 1112–1137. MR3649625 https://doi.org/10.1080/17415977.2016.1222528

ZHANG, Y., LIN, G.-L., FORSSÉN, P., GULLIKSSON, M., FORNSTEDT, T. and CHENG, X.-L. (2016). A regularization method for the reconstruction of adsorption isotherms in liquid chromatography. *Inverse Probl.* **32** 105005, 24. MR3627029 https://doi.org/10.1088/0266-5611/32/10/105005

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3500–3525
https://doi.org/10.1214/23-AOAS1773

ACCOUNTING FOR SEASONALITY IN EXTREME SEA-LEVEL ESTIMATION

BY ELEANOR D'ARCY[1,a], JONATHAN A. TAWN[1,b] AMÉLIE JOLY[2,c] AND DAFNI E. SIFNIOTI[2,d]

[1]*STOR-i Centre for Doctoral Training, Department of Mathematics and Statistics, Lancaster University,* [a]*e.darcy@lancaster.ac.uk,* [b]*j.tawn@lancaster.ac.uk*

[2]*EDF Energy R&D UK Centre,* [c]*amelie.joly@edf.fr,* [d]*dafni.sifnioti@edfenergy.com*

Reliable estimates of sea-level return-levels are crucial for coastal flooding risk assessments and for coastal flood defence design. We describe a novel method for estimating extreme sea-levels that is the first to capture seasonality, interannual variations and longer term changes. We use a joint probabilities method, with skew-surge and peak-tide as two sea-level components. The tidal regime is predictable, but skew-surges are stochastic. We present a statistical model for skew-surges, where the main body of the distribution is modelled empirically while a nonstationary generalised Pareto distribution (GPD) is used for the upper tail. We capture within-year seasonality by introducing a daily covariate to the GPD model and allowing the distribution of peak-tide to change over months and years. Skew-surge-peak-tide dependence is accounted for, via a tidal covariate, in the GPD model, and we adjust for skew-surge temporal dependence through the subasymptotic extremal index. We incorporate spatial prior information in our GPD model to reduce the uncertainty associated with the highest return-level estimates. Our results are an improvement on current return-level estimates, with previous methods typically underestimating. We illustrate our method at four U.K. tide gauges.

REFERENCES

ASADI, P., ENGELKE, S. and DAVISON, A. C. (2018). Optimal regionalization of extreme value distributions for flood estimation. *J. Hydrol.* **556** 182–193.

BARANES, H., WOODRUFF, J., TALKE, S., KOPP, R., RAY, R. and DECONTO, R. (2020). Tidally driven interannual variation in extreme sea-level frequencies in the Gulf of Maine. *J. Geophys. Res., Oceans* **125** e2020JC016291.

BASHTANNYK, D. M. and HYNDMAN, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.* **36** 279–298. MR1836204 https://doi.org/10.1016/S0167-9473(00)00046-3

BATSTONE, C., LAWLESS, M., TAWN, J. A., HORSBURGH, K., BLACKMAN, D., MCMILLAN, A., WORTH, D., LAEGER, S. and HUNT, T. (2013). A UK best-practice approach for extreme sea-level analysis along complex topographic coastlines. *Ocean Eng.* **71** 28–39.

BERNARDARA, P., ANDREEWSKY, M. and BENOIT, M. (2011). Application of regional frequency analysis to the estimation of extreme storm surges. *J. Geophys. Res., Oceans* **116**.

CARTER, D. and CHALLENOR, P. (1981). Estimating return values of environmental parameters. *Q. J. R. Meteorol. Soc.* **107** 259–266.

CHAVEZ-DEMOULIN, V. and DAVISON, A. C. (2005). Generalized additive modelling of sample extremes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 207–222. MR2134607 https://doi.org/10.1111/j.1467-9876.2005.00479.x

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics*. Springer London, Ltd., London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

COLES, S. G., HEFFERNAN, J. and TAWN, J. A. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.

COLES, S. G. and TAWN, J. A. (1990). Statistics of coastal flood prevention. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **332** 457–476.

Key words and phrases. Extreme sea-levels, generalised Pareto distribution, joint probabilities method, nonstationarity, skew-surge.

COLES, S. G., TAWN, J. A. and SMITH, R. L. (1994). A seasonal Markov model for extremely low temperatures. *Environmetrics* **5** 221–239.
D'ARCY, E., TAWN, J. A. and SIFNIOTI, D. E. (2022). Accounting for climate change in extreme sea-level estimation. *Water* **14** 2956.
D'ARCY, E., TAWN, J. A. JOLY, A. and SIFNIOTI, D. E. (2023). Supplement to "Accounting for seasonality in extreme sea-level estimation." https://doi.org/10.1214/23-AOAS1773SUPPA, https://doi.org/10.1214/23-AOAS1773SUPPB
DAVISON, A., HUSER, R. and THIBAUD, E. (2019). Spatial extremes. In *Handbook of Environmental and Ecological Statistics*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 711–744. CRC Press, Boca Raton, FL. MR3889918
DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. MR1086795
DIXON, M. and TAWN, J. A. (1994). Extreme sea-levels: Modelling interaction between tide and surge. *Statistics for the Environment* 2: *Water Related Issues* 221–232.
DIXON, M. J. and TAWN, J. A. (1999). The effect of non-stationarity on extreme sea-level estimation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 135–151.
DIXON, M. J., TAWN, J. A. and VASSIE, J. M. (1998). Spatial modelling of extreme sea-levels. *Environmetrics* **9** 283–301.
EASTOE, E. F. and TAWN, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 25–45. MR2662232 https://doi.org/10.1111/j.1467-9876.2008.00638.x
EGBERT, G. D. and RAY, R. D. (2017). Tidal prediction. *J. Mar. Res.* **75** 189–237.
ENVIRONMENT AGENCY (2018). Coastal Flood Boundary Conditions for the UK: Update 2018. Technical summary report. Available at https://environment.data.gov.uk/dataset/6e856bda-0ca9-404f-93d7-566a2378a7a8.
FAWCETT, L. and WALSHAW, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics* **18** 173–188. MR2345653 https://doi.org/10.1002/env.810
FERRO, C. A. T. and SEGERS, J. (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 545–556. MR1983763 https://doi.org/10.1111/1467-9868.00401
GRAFF, J. (1978). Concerning the recurrence of abnormal sea-levels. *Coast. Eng.* **2** 177–187.
HAIGH, I. D., NICHOLLS, R. and WELLS, N. (2010). A comparison of the main methods for estimating probabilities of extreme still water levels. *Coast. Eng.* **57** 838–849.
HOSKING, J. R. M. and WALLIS, J. R. (1997). *Regional Frequency Analysis*: *An Approach Based on L-Moments*. Cambridge Univ. Press, Cambridge; New York.
HOWARD, T. and WILLIAMS, S. D. P. (2021). Towards using state-of-the-art climate models to help constrain estimates of unprecedented UK storm surges. *Nat. Hazards Earth Syst. Sci.* **21** 3693–3712.
JONATHAN, P., RANDELL, D., WU, Y. and EWANS, K. (2014). Return level estimation from nonstationary spatial data exhibiting multidimensional covariate effects. *Ocean Eng.* **88** 520–532.
LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. *Springer Series in Statistics*. Springer, New York-Berlin. MR0691492
LEDFORD, A. W. and TAWN, J. A. (2003). Diagnostics for dependence within time series extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 521–543. MR1983762 https://doi.org/10.1111/1467-9868.00400
MENÉNDEZ, M. and WOODWORTH, P. L. (2010). Changes in extreme high water levels based on a quasi-global tide-gauge data set. *J. Geophys. Res., Oceans* **115**.
NOC (2021). National Tidal and Sea-Level Facility. Available at https://www.ntslf.org/.
NORTHROP, P. J., JONATHAN, P. and RANDELL, D. (2016). Threshold modeling of nonstationary extremes. In *Extreme Value Modeling and Risk Analysis* 87–108. CRC Press, Boca Raton, FL. MR3644309
POLITIS, D. N. and ROMANO, J. P. (1994). The stationary bootstrap. *J. Amer. Statist. Assoc.* **89** 1303–1313. MR1310224
PRANDLE, D. and WOLF, J. (1978). Surge-tide interaction in the southern North Sea. *Elsevier Oceanography Series* **23** 161–185.
PUGH, D. and VASSIE, J. (1978). Extreme sea-levels from tide and surge probability. *Coast. Eng.* **16** 911–930.
PUGH, D. and WOODWORTH, P. (2014). *Sea-Level Science*: *Understanding Tides*, *Surges*, *Tsunamis and Mean Sea-Level Changes*. Cambridge Univ. Press, Cambridge.
ROBINSON, M. E. and TAWN, J. A. (1997). Statistics for extreme sea currents. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **46** 183–205.
RODRÍGUEZ ENRÍQUEZ, A., WAHL, T., BARANES, H., TALKE, S. A., ORTON, P. M., BOOTH, J. F. and HAIGH, I. D. (2021). Predictable changes in extreme sea-levels and coastal flood risk due to nodal and perigean astronomical tidal cycles. *Earth and Space Science Open Archive*.
ROHRBECK, C. and TAWN, J. A. (2021). Bayesian spatial clustering of extremal behavior for hydrological variables. *J. Comput. Graph. Statist.* **30** 91–105. MR4235967 https://doi.org/10.1080/10618600.2020.1777139

ROOTZÉN, H. and KATZ, R. W. (2013). Design life level: Quantifying risk in a changing climate. *Water Resour. Res*. **49** 5964–5972.

SHARKEY, P. and WINTER, H. C. (2019). A Bayesian spatial hierarchical model for extreme precipitation in Great Britain. *Environmetrics* **30** e2529. MR3908107 https://doi.org/10.1002/env.2529

SMITH, R. L. and WEISSMAN, I. (1994). Estimating the extremal index. *J. Roy. Statist. Soc. Ser. B* **56** 515–528. MR1278224

TAWN, J. A. and VASSIE, J. M. and GUMBEL, E. J. (1989). Extreme sea-levels: The joint probabilities method revisited and revised. *Proceedings of the Institution of Civil Engineers* **87** 429–442.

TAWN, J. A. (1988). An extreme-value theory model for dependent observations. *J. Hydrol*. **101** 227–250.

TAWN, J. A. (1992). Estimating probabilities of extreme sea-levels. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **41** 77–93.

WADEY, M. P., HAIGH, I. D., NICHOLLS, R. J., BROWN, J. M., HORSBURGH, K., CARROLL, B., GALLOP, S. L., MASON, T., BRADSHAW, E. et al. (2015). A comparison of the 31 January–1 February 1953 and 5–6 December 2013 coastal flood events around the UK. *Front. Mar. Sci*. **2** 84.

WILLIAMS, J., HORSBURGH, K. J., WILLIAMS, J. A. and PROCTOR, R. N. (2016). Tide and skew-surge independence: New insights for flood risk. *Geophys. Res. Lett*. **43** 6410–6417.

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3526–3549
https://doi.org/10.1214/23-AOAS1774

ASSOCIATION AND CAUSATION: ATTRIBUTES AND EFFECTS OF JUDGES IN EQUAL EMPLOYMENT OPPORTUNITY COMMISSION LITIGATION OUTCOMES

BY MICHAEL E. SOBEL[1,a], GREGORY J. WAWRO[2,b] AND SEAN FARHANG[3,c]

[1]*Department of Statistics, Columbia University,* [a]*mes105@columbia.edu*
[2]*Department of Political Science, Columbia University,* [b]*gjw10@columbia.edu*
[3]*School of Law, University of California, Berkeley,* [c]*farhang@berkeley.edu*

A large literature on judicial decision making asks if judges with different features of an attribute (e.g, sex, race) adjudicate cases differently. Researchers estimate models for case outcomes, interpreting coefficients associated with attributes as effects. But attributes are not treatments. While these coefficients indicate how judges with different features adjudicate the different cases they are assigned, ideally, different judges should be compared on a common set of cases. We construct a general methodology for making such comparisons, using it to study whether monetary relief in discrimination cases brought by the Equal Employment Opportunity Commission differs by judges' race. For all federal judges (treatments) eligible to hear a case (unit), we define potential outcomes, using unit treatment effects between judges with different features to define a unit feature comparison (UFC), then using these to define new population estimands: the average (AFC) and quantile (QFC) feature comparisons. We estimate these quantities by combining observed case outcomes with missing potential outcomes imputed from the posterior predictive distribution of a two-part Bayesian hierarchical model. A case initially assigned to a non-white or African American judge is more likely to result in monetary relief than were that case initially assigned to an eligible white or non-African American judge. For the amount of relief, the 95% posterior interval for the AFC covers 0, while the upper endpoint of the 95% posterior interval for the median QFC is negative.

REFERENCES

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

ASHENFELTER, O., EISENBERG, T. and SCHWAB, S. J. (1995). Politics and the judiciary: The influence of judicial background on case outcomes. *J. Leg. Stud.* **24** 257–281.

BELSON, W. A. (1956). A technique for studying the effects of a television broadcast. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **5** 195–202.

BOYD, C. L. (2016). Representation on the courts? The effects of trial judges' sex and race. *Polit. Res. Q.* **69** 788–799.

BOYD, C. L., EPSTEIN, L. and MARTIN, A. D. (2010). Untangling the causal effects of sex on judging. *Amer. J. Polit. Sci.* **54** 389–411.

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**.

COHEN, A. and YANG, C. (2018). Judicial Politics and Sentencing Decisions Technical Report National Bureau of Economic Research. NBER Working Paper No. w24615.

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; CRC Press, London. MR0095561

DAVIS, G. C. JR and KUTNER, M. H. (1976). The lagged normal family of probability density functions applied to indicator-dilution curves. *Biometrics* **32** 669–675.

Key words and phrases. Causal inference, potential outcomes, unit feature comparison, average feature comparison, quantile feature comparison, U.S. Federal Courts, Bayesian hierarchical model.

DUAN, N., MANNING, W. G., MORRIS, C. N. and NEWHOUSE, J. P. (1983). A comparison of alternative models for the demand for medical care. *J. Bus. Econom. Statist.* **1** 115–126.

EISENBERG, T., EISENBERG, T., WELLS, M. T. and ZHANG, M. (2015). Addressing the zeros problem: Regression models for outcomes with a large proportion of zeros, with an application to trial outcomes. *J. Empir. Leg. Stud.* **12** 161–186.

EMERSON, J. W., SELTZER, M. and LIN, D. (2009). Assessing judging bias: An example from the 2000 Olympic Games. *Amer. Statist.* **63** 124–131. MR2750072 https://doi.org/10.1198/tast.2009.0026

FARHANG, S. and WAWRO, G. (2004). Institutional dynamics on the U.S. court of appeals: Minority representation under panel decision making. *J. Law Econ. Organ.* **20** 299–330.

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

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

GEORGE, T. E. and YOON, A. H. (2017). Measuring justice in state courts: The demographics of the state judiciary. *Vanderbilt Law Rev.* **70** 1887–1910.

GOLDMAN, S. (1966). Voting behavior on the United States Courts of Appeals, 1961–1964. *Amer. Polit. Sci. Rev.* **60** 374–383.

GREINER, D. J. and RUBIN, D. B. (2011). Causal effects of perceived immutable characteristics. *Rev. Econ. Stat.* **93** 775–785.

HANEY, S. (2011). *Practical Application and Properties of the Exponentially Modified Gaussian* (*EMG*) *Distribution*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Drexel University. MR2941787

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

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

JOHNSON, B. D. (2014). Judges on trial: A reexamination of judicial race and gender effects across modes of conviction. *Crim. Justice Policy Rev.* **25** 159–184.

KASTELLEC, J. P. (2013). Racial diversity and judicial influence on appellate courts. *Amer. J. Polit. Sci.* **57** 167–83.

KULIK, C. T., PERRY, E. L. and PEPPER, M. B. (2003). Here comes the judge: The influence of judge personal characteristics on federal sexual harassment case outcomes. *Law Hum. Behav.* **27** 69–86. https://doi.org/10.1023/a:1021678912133

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

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438 https://doi.org/10.1198/016214501753168389

RAMJI-NOGALES, J., SCHOENHOLTZ, A. I. and SCHRAG, P. G. (2007). Refugee roulette: Disparities in asylum adjudication. *Stanf. Law Rev.* **60** 295–412.

ROBERTSON, C., BAUGHMAN, S. B. and WRIGHT, M. S. (2019). Race and class: A randomized experiment with prosecutors. *J. Empir. Leg. Stud.* **16** 807–847.

ROOT, D., FALESCHINI, J. and OYENUBI, G. (2019). Building a More Inclusive Federal Judiciary. Center for American Progress.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SEGAL, J. A. (2000). Representative decision making on the federal bench: Clinton's district court appointees. *Polit. Res. Q.* **53** 137–150.

SUNSTEIN, C. R., SCHKADE, D., ELLMAN, L. M. and SAWICKI, A. (2006). *Are Judges Political?: An Empirical Analysis of the Federal Judiciary*. Brookings Institution Press, Washington, D.C.

TANNER, M. A. and YOUNG, M. A. (1985). Modeling agreement among raters. *J. Amer. Statist. Assoc.* **80** 175–180.

VANBELLE, S. and ALBERT, A. (2009). Agreement between two independent groups of raters. *Psychometrika* **74** 477–491. MR2551672 https://doi.org/10.1007/s11336-009-9116-1

VANDERWEELE, T. J. and HERNÁN, M. A. (2013). Causal inference under multiple versions of treatment. *J. Causal Inference* **1** 1–20. MR4289399 https://doi.org/10.1515/jci-2012-0002

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

WALKER, T. G. and BARROW, D. J. (1985). The diversification of the federal bench: Policy and process ramifications. *J. Polit.* **47** 596–617.

WEINBERG, J. D. and NIELSEN, L. B. (2012). Examining empathy: Discrimination, experience, and judicial decisionmaking. *South. Calif. Law Rev.* **85** 313–52.

ZHANG, M., STRAWDERMAN, R. L., COWEN, M. E. and WELLS, M. T. (2006). Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *J. Amer. Statist. Assoc.* **101** 934–945. MR2324094 https://doi.org/10.1198/016214505000001429

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3550–3569
https://doi.org/10.1214/23-AOAS1775

DEBIASED LASSO FOR STRATIFIED COX MODELS WITH APPLICATION TO THE NATIONAL KIDNEY TRANSPLANT DATA

BY LU XIA[1,a] BIN NAN[2,b] AND YI LI[3,c]

[1]*Department of Biostatistics, University of Washington,* [a]*xialu@uw.edu*
[2]*Department of Statistics, University of California, Irvine,* [b]*nanb@uci.edu*
[3]*Department of Biostatistics, University of Michigan,* [c]*yili@umich.edu*

The Scientific Registry of Transplant Recipients (SRTR) system has become a rich resource for understanding the complex mechanisms of graft failure after kidney transplant, a crucial step for allocating organs effectively and implementing appropriate care. As transplant centers that treated patients might strongly confound graft failures, Cox models stratified by centers can eliminate their confounding effects. Also, since recipient age is a proven nonmodifiable risk factor, a common practice is to fit models separately by recipient age groups. The moderate sample sizes, relative to the number of covariates, in some age groups may lead to biased maximum stratified partial likelihood estimates and unreliable confidence intervals, even when samples still outnumber covariates. To draw reliable inference on a comprehensive list of risk factors measured from both donors and recipients in SRTR, we propose a debiased lasso approach via quadratic programming for fitting stratified Cox models. We establish asymptotic properties and verify via simulations that our method produces consistent estimates and confidence intervals with nominal coverage probabilities. Accounting for nearly 100 confounders in SRTR, the debiased method detects that the graft failure hazard nonlinearly increases with donor's age among all recipient age groups and that organs from older donors more adversely impact the younger recipients. Our method also delineates the associations between graft failure and many risk factors such as recipients' primary diagnoses (e.g., polycystic disease, glomerular disease, and diabetes) and donor-recipient mismatches for human leukocyte antigen loci across recipient age groups. These results may inform the refinement of donor-recipient matching criteria for stakeholders.

REFERENCES

ALEXANDER, J. W., BENNETT, L. E. and BREEN, T. J. (1994). Effect of donor age on outcome of kidney transplantation. A two-year analysis of transplants reported to the united network for organ sharing registry. *Transplantation* **57** 871–876.

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

BAKER, R. J., MARK, P. B., PATEL, R. K., STEVENS, K. K. and PALMER, N. (2017). Renal association clinical practice guideline in post-operative care in the kidney transplant recipient. *BMC Nephrol.* **18** 174. https://doi.org/10.1186/s12882-017-0553-2

BASTANI, B. (2015). The worsening transplant organ shortage in USA; desperate times demand innovative solutions. *J. Nephropathol.* **4** 105–109. https://doi.org/10.12860/jnp.2015.20

CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. MR3476606 https://doi.org/10.1214/13-AOS1171

DAYOUB, J. C., CORTESE, F., ANŽIČ, A., GRUM, T. and DE MAGALHÃES, J. P. (2018). The effects of donor age on organ transplants: A review and implications for aging research. *Exp. Gerontol.* **110** 230–240. https://doi.org/10.1016/j.exger.2018.06.019

DICKINSON, D. M., ARRINGTON, C. J., FANT, G., LEVINE, G. N., SCHAUBEL, D. E., PRUETT, T. L., ROBERTS, M. S. and WOLFE, R. A. (2008). SRTR program-specific reports on outcomes: A guide for the new reader. *Am. J. Transplant.* **8** 1012–1026.

Key words and phrases. Confidence intervals, diverging number of covariates, end-stage renal disease, graft failure free survival, statistical inference.

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 https://doi.org/10.1214/aos/1015362185

FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 https://doi.org/10.1111/rssb.12224

FARAVARDEH, A., EICKHOFF, M., JACKSON, S., SPONG, R., KUKLA, A., ISSA, N., MATAS, A. J. and IBRAHIM, H. N. (2013). Predictors of graft failure and death in elderly kidney transplant recipients. *Transplantation* **96** 1089–1096. https://doi.org/10.1097/TP.0b013e3182a688e5

FEI, Z. and LI, Y. (2021). Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. *J. Mach. Learn. Res.* **22** 58. MR4253751

FERRARI, P., LIM, W., DENT, H. and MCDONALD, S. P. (2011). Effect of donor-recipient age difference on graft function and survival in live-donor kidney transplantation. *Nephrol. Dial. Transplant.* **26** 702–708.

HAMIDI, O., POOROLAJAL, J., FARHADIAN, M. and TAPAK, L. (2016). Identifying important risk factors for survival in kidney graft failure patients using random survival forests. *Iran. J. Public Health* **45** 27–33.

HE, K., ZHU, J., KANG, J. and LI, Y. (2022). Stratified Cox models with time-varying effects for national kidney transplant patients: A new blockwise steepest ascent method. *Biometrics* **78** 1221–1232. MR4493519

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124 https://doi.org/10.1006/jmva.1999.1873

HUANG, J., SUN, T., YING, Z., YU, Y. and ZHANG, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41** 1142–1165. MR3113806 https://doi.org/10.1214/13-AOS1098

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

JU, A., CHOW, B. Y., RALPH, A. F., HOWELL, M., JOSEPHSON, M. A., AHN, C., BUTT, Z., DOBBELS, F., FOWLER, K. et al. (2019). Patient-reported outcome measures for life participation in kidney transplantation: A systematic review. *Am. J. Transplant.* **19** 2306–2317.

KABORÉ, R., COUCHOUD, C., MACHER, M.-A., SALOMON, R., RANCHIN, B., LAHOCHE, A., ROUSSEY-KESLER, G., GARAIX, F., DECRAMER, S. et al. (2017). Age-dependent risk of graft failure in young kidney transplant recipients. *Transplantation* **101** 1327–1335.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR1924807 https://doi.org/10.1002/9781118032985

KARIM, A., FARRUGIA, D., CHESHIRE, J., MAHBOOB, S., BEGAJ, I., RAY, D. and SHARIF, A. (2014). Recipient age and risk for mortality after kidney transplantation in England. *Transplantation* **97** 832–838.

KASISKE, B. L., ISRANI, A. K., SNYDER, J. J., SKEANS, M. A. and PATIENT OUTCOMES IN RENAL TRANSPLANTATION (PORT) INVESTIGATORS (2011). The relationship between kidney function and long-term graft survival after kidney transplant. *Am. J. Kidney Dis.* **57** 466–475. https://doi.org/10.1053/j.ajkd.2010.10.054

KASISKE, B. L. and SNYDER, J. (2002). Matching older kidneys with older patients does not improve allograft survival. *J. Am. Soc. Nephrol.* **13** 1067–1072.

KEITH, D. S., DEMATTOS, A., GOLCONDA, M., PRATHER, J. and NORMAN, D. (2004). Effect of donor recipient age match on survival after first deceased donor renal transplantation. *J. Am. Soc. Nephrol.* **15** 1086–1091. https://doi.org/10.1097/01.asn.0000119572.02053.f2

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787 https://doi.org/10.1214/aos/1015957397

KONG, S. and NAN, B. (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via lasso. *Statist. Sinica* **24** 25–42. MR3184591

KONG, S., YU, Z., ZHANG, X. and CHENG, G. (2021). High-dimensional robust inference for Cox regression models using desparsified Lasso. *Scand. J. Stat.* **48** 1068–1095. MR4303569 https://doi.org/10.1111/sjos.12543

KOSTRO, J. Z., HELLMANN, A., KOBIELA, J., SKÓRA, I., LICHODZIEJEWSKA-NIEMIERKO, M., DĘBSKA-ŚLIZIEŃ, A. and ŚLEDZIŃSKI, Z. (2016). Quality of life after kidney transplantation: A prospective study. *Transplant. Proc.* **48** 50–54. https://doi.org/10.1016/j.transproceed.2015.10.058

KOVESDY, C. P., PARK, J. C. and KALANTAR-ZADEH, K. (2010). Glycemic control and burnt-out diabetes in ESRD. In *Semin. Dial.* **23** 148–156. Wiley, New York.

LEGENDRE, C., CANAUD, G. and MARTINEZ, F. (2014). Factors influencing long-term outcome after kidney transplantation. *Transpl. Int.* **27** 19–27. https://doi.org/10.1111/tri.12217

LIM, W. H., CHANG, S., CHADBAN, S., CAMPBELL, S., DENT, H., RUSS, G. R. and MCDONALD, S. P. (2010). Donor-recipient age matching improves years of graft function in deceased-donor kidney transplantation. *Nephrol. Dial. Transplant.* **25** 3082–3089.

MORALES, J. M., MARCÉN, R., DEL CASTILLO, D., ANDRES, A., GONZALEZ-MOLINA, M., OPPENHEIMER, F., SERÓN, D., GIL-VERNET, S., LAMPREAVE, I. et al. (2012). Risk factors for graft loss and mortality after

renal transplantation according to recipient age: A prospective multicentre study. *Nephrol. Dial. Transplant.* **27** iv39–iv46. https://doi.org/10.1093/ndt/gfs544

RAO, P. S. and OJO, A. (2009). The alphabet soup of kidney transplantation: SCD, DCD, ECD–fundamentals for the practicing nephrologist. *Clin. J. Amer. Soc. Nephrol.* **4** 1827–1831. https://doi.org/10.2215/CJN.02270409

RAO, P. S., SCHAUBEL, D. E., GUIDINGER, M. K., ANDREONI, K. A., WOLFE, R. A., MERION, R. M., PORT, F. K. and SUNG, R. S. (2009). A comprehensive risk quantification score for deceased donor kidneys: The kidney donor risk index. *Transplantation* **88** 231–236. https://doi.org/10.1097/TP.0b013e3181ac620b

RODGER, R. S. C. (2012). Approach to the management of end-stage renal disease. *Clin. Med.* **12** 472–475.

SARAN, R., ROBINSON, B., ABBOTT, K. C., BRAGG-GRESHAM, J., CHEN, X., GIPSON, D., GU, H., HIRTH, R. A., HUTTON, D. et al. (2020). US renal data system 2019 annual data report: Epidemiology of kidney disease in the United States. *Am. J. Kidney Dis.* **75** Svi–Svii.

SHI, X., LV, J., HAN, W., ZHONG, X., XIE, X., SU, B. and DING, J. (2018). What is the impact of human leukocyte antigen mismatching on graft survival and mortality in renal transplantation? A meta-analysis of 23 cohort studies involving 486,608 recipients. *BMC Nephrol.* **19** 116.

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13. https://doi.org/10.18637/jss.v039.i05

SMITH, J., BIGGINS, S., HASELBY, D., KIM, W., WEDD, J., LAMB, K., THOMPSON, B., SEGEV, D., GUSTAFSON, S. et al. (2012). Kidney, pancreas and liver allocation and distribution in the United States. *Am. J. Transplant.* **12** 3191–3212.

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

VAIDA, F. and XU, R. (2000). Proportional hazards model with random effects. *Stat. Med.* **19** 3309–3324.

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

VEROUX, M., GROSSO, G., CORONA, D., MISTRETTA, A., GIAQUINTA, A., GIUFFRIDA, G., SINAGRA, N. and VEROUX, P. (2012). Age is an important predictor of kidney transplantation outcome. *Nephrol. Dial. Transplant.* **27** 1663–1671. https://doi.org/10.1093/ndt/gfr524

WANG, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.* **39** 389–417. MR2797851 https://doi.org/10.1214/10-AOS846

WOLFE, R. A. (1991). Survival analysis methods for the end-stage renal disease (ESRD) program of medicare. In *Kidney Failure and the Federal Government* 353–400 (R. A. Rettig and N. G. Levinsky, eds.) National Academies Press, Washington, DC.

WOLFE, R. A., ASHBY, V. B., MILFORD, E. L., OJO, A. O., ETTENGER, R. E., AGODOA, L. Y., HELD, P. J. and PORT, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *N. Engl. J. Med.* **341** 1725–1730. https://doi.org/10.1056/NEJM199912023412303

XIA, L., NAN, B. and LI, Y. (2023a). Statistical inference for Cox proportional hazards models with a diverging number of covariates. *Scand. J. Stat.* **50** 550–571. MR4599924 https://doi.org/10.1111/sjos.12595

XIA, L., NAN, B. and LI, Y. (2023). Supplement to "Debiased lasso for stratified Cox models with application to the national kidney transplant data." https://doi.org/10.1214/23-AOAS1775SUPPA, https://doi.org/10.1214/23-AOAS1775SUPPB

YU, Y., BRADIC, J. and SAMWORTH, R. J. (2021). Confidence intervals for high-dimensional Cox models. *Statist. Sinica* **31** 243–267.

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

The Annals of Applied Statistics
2023, Vol. 17, No. 4, 3570–3588
https://doi.org/10.1214/23-AOAS1776

CONTINUOUS-TIME MODELLING OF BEHAVIOURAL RESPONSES IN ANIMAL MOVEMENT

BY THÉO MICHELOT[1,a], RICHARD GLENNIE[2,b], LEN THOMAS[2,c], NICOLA QUICK[3,e] AND CATRIONA M. HARRIS[2,d]

[1]*Department of Mathematics and Statistics, Dalhousie University,* [a]*theo.michelot@dal.ca*
[2]*Centre for Research into Ecological and Environmental Modelling, University of St Andrews,* [b]*research@glennies.co.uk,* [c]*len.thomas@st-andrews.ac.uk,* [d]*catriona.harris@st-andrews.ac.uk*
[3]*Nicholas School of the Environment, Duke University,* [e]*nicola.quick@plymouth.ac.uk*

There is great interest in ecology to understand how wild animals are affected by anthropogenic disturbances, such as sounds. For example, behavioural response studies are an important approach to quantify the impact of naval activity on marine mammals. Controlled exposure experiments are undertaken where the behaviour of animals is quantified before, during, and after exposure to a controlled sound source, often using telemetry tags (e.g., accelerometers or satellite trackers). Statistical modelling is required to formally compare patterns before and after exposure, to quantify deviations from baseline behaviour. We propose varying-coefficient stochastic differential equations (SDEs) as a flexible framework to model such data with two components: (1) time-varying baseline dynamics, modelled with nonparametric or random effects of time-varying covariates, and (2) a nonparametric response model, which captures deviations from baseline. SDEs are specified in continuous time, which makes it straightforward to analyse data collected at irregular time intervals, a common situation for animal tracking studies. We describe how the model can be embedded into a state-space modelling framework to account for measurement error. We present inferential methods for model fitting, model checking, and uncertainty quantification (including on the response model). We apply this approach to two behavioural response study data sets on beaked whales: a satellite track and high-resolution depth data. Our results suggest that the whales' horizontal movement and vertical diving behaviour changed after exposure to the sound source, and future work should evaluate the severity and possible consequences of these responses. These two very different examples showcase the versatility of varying-coefficient SDEs to measure changes in behaviour, and we discuss implications of disturbances for the whales' energetic balance.

REFERENCES

ANDERSON-SPRECHER, R. and LEDOLTER, J. (1991). State-space analysis of wildlife telemetry data. *J. Amer. Statist. Assoc.* **86** 596–602.

ANDREWS, R. D., PITMAN, R. L. and BALLANCE, L. T. (2008). Satellite tracking reveals distinct movement patterns for Type B and Type C killer whales in the southern Ross Sea, Antarctica. *Polar Biol.* **31** 1461–1468.

CIOFFI, W. R., QUICK, N. J., SWAIM, Z. T., FOLEY, H. J., WAPLES, D. M., WEBSTER, D. L., BAIRD, R. W., SOUTHALL, B. L., NOWACEK, D. P. et al. (2022). Trade-offs in telemetry tag programming for deep-diving cetaceans: Data longevity, resolution, and continuity. *BioRxiv*.

DERUITER, S. L., LANGROCK, R., SKIRBUTAS, T., GOLDBOGEN, J. A., CALAMBOKIDIS, J., FRIEDLAENDER, A. S. and SOUTHALL, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *Ann. Appl. Stat.* **11** 362–392. MR3634328 https://doi.org/10.1214/16-AOAS1008

DERUITER, S. L., SOUTHALL, B. L., CALAMBOKIDIS, J., ZIMMER, W. M., SADYKOVA, D., FALCONE, E. A., FRIEDLAENDER, A. S., JOSEPH, J. E., MORETTI, D. et al. (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biol. Lett.* **9** 20130223.

Key words and phrases. Stochastic differential equation, diffusion process, behavioural response study, beaked whale.

DUNN, J. E. and GIPSON, P. S. (1977). Analysis of radio telemetry data in studies of home range. *Biometrics* 85–101.

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

ELERIAN, O., CHIB, S. and SHEPHARD, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69** 959–993. MR1839375 https://doi.org/10.1111/1468-0262.00226

FOLEY, H. J., PACIFICI, K., BAIRD, R. W., WEBSTER, D. L., SWAIM, Z. T. and READ, A. J. (2021). Residency and movement patterns of Cuvier's beaked whales Ziphius cavirostris off Cape Hatteras, North Carolina, USA. *Mar. Ecol. Prog. Ser.* **660** 203–216.

GURARIE, E., CAGNACCI, F., PETERS, W., FLEMING, C. H., CALABRESE, J. M., MUELLER, T. and FAGAN, W. F. (2017). A framework for modelling range shifts and migrations: Asking when, whither, whether and will it return. *J. Anim. Ecol.* **86** 943–959. https://doi.org/10.1111/1365-2656.12674

JOHNSON, D. S. and LONDON, J. M. (2018). crawl: An R package for fitting continuous-time correlated random walk models to animal movement data.

JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215. https://doi.org/10.1890/07-1032.1

JOHNSON, M. P. and TYACK, P. L. (2003). A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE J. Oceanic Eng.* **28** 3–12.

JONES-TODD, C. M., PIROTTA, E., DURBAN, J. W., CLARIDGE, D. E., BAIRD, R. W., FALCONE, E. A., SCHORR, G. S., WATWOOD, S. and THOMAS, L. (2022). Discrete-space continuous-time models of marine mammal exposure to Navy sonar. *Ecol. Appl.* **32** e02475.

JONSEN, I. D., MYERS, R. A. and MILLS FLEMMING, J. (2003). Meta-analysis of animal movement using state-space models. *Ecology* **84** 3055–3063.

KRISTENSEN, K., NIELSEN, A., BERG, C. W., SKAUG, H. and BELL, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *J. Stat. Softw.* **70** 1–21.

MARRA, G. and WOOD, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scand. J. Stat.* **39** 53–74. MR2896791 https://doi.org/10.1111/j.1467-9469.2011.00760.x

MCCLINTOCK, B. T., LONDON, J. M., CAMERON, M. F. and BOVENG, P. L. (2015). Modelling animal movement using the argos satellite telemetry location error ellipse. *Methods Ecol. Evol.* **6** 266–277.

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969 https://doi.org/10.1214/aos/1176325622

MICHELOT, T., GLENNIE, R., HARRIS, C. and THOMAS, L. (2021). Varying-coefficient stochastic differential equations with applications in ecology. *J. Agric. Biol. Environ. Stat.* **26** 446–463. MR4292797 https://doi.org/10.1007/s13253-021-00450-6

MICHELOT, T., GLENNIE, R., THOMAS, L., QUICK, N. and HARRIS, C. M. (2023a). Code and data for "Continuous-time modelling of behavioural responses in animal movement." https://doi.org/10.1214/23-AOAS1776SUPPA

MICHELOT, T., GLENNIE, R., THOMAS, L., QUICK, N. and HARRIS, C. M. (2023b). Supplementary materials for "Continuous-time modelling of behavioural responses in animal movement." https://doi.org/10.1214/23-AOAS1776SUPPB

MICHELOT, T., GLOAGUEN, P., BLACKWELL, P. G. and ÉTIENNE, M.-P. (2019). The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods Ecol. Evol.* **10** 1894–1907.

MILLER, D. L. (2019). Bayesian views of generalized additive modelling. ArXiv preprint. Available at arXiv:1902.01330.

MILLER, P. J., KVADSHEIM, P. H., LAM, F.-P. A., WENSVEEN, P. J., ANTUNES, R., ALVES, A. C., VISSER, F., KLEIVANE, L., TYACK, P. L. et al. (2012). The severity of behavioral changes observed during experimental exposures of killer (Orcinus orca), long-finned pilot (Globicephala melas), and sperm (Physeter macrocephalus) whales to naval sonar. *Aquat. Mamm.* **38** 362.

MILNER, J. E., BLACKWELL, P. G. and NIU, M. (2021). Modelling and inference for the movement of interacting animals. *Methods Ecol. Evol.* **12** 54–69.

NIU, M., BLACKWELL, P. G. and SKARIN, A. (2016). Modeling interdependent animal movement in continuous time. *Biometrics* **72** 315–324. MR3515758 https://doi.org/10.1111/biom.12454

POZDNYAKOV, V., MEYER, T., WANG, Y.-B. and YAN, J. (2014). On modeling animal movements using Brownian motion with measurement error. *Ecology* **95** 247–253.

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

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 https://doi.org/10.1017/CBO9780511755453

SHEARER, J. M., QUICK, N. J., CIOFFI, W. R., BAIRD, R. W., WEBSTER, D. L., FOLEY, H. J., SWAIM, Z. T., WAPLES, D. M., BELL, J. T. et al. (2019). Diving behaviour of Cuvier's beaked whales (Ziphius cavirostris) off Cape Hatteras, North Carolina. *R. Soc. Open Sci.* **6** 181728. https://doi.org/10.1098/rsos.181728

SOUTHALL, B. L., BOWERS, M., CIOFFI, W., FOLEY, H., HARRIS, C., JOSEPH, J., QUICK, N., MARGOLINA, T., NOWACEK, D. et al. (2020). Atlantic Behavioral Response Study (BRS): 2019 Annual Progress Report. Project report. Prepared for U.S. Fleet Forces Command. Submitted to Naval Facilities Engineering Command Atlantic, Norfolk, Virginia, under Contract No. N62470-15-D-8006, Task Order 19F4029, issued to HDR Inc., Virginia Beach, Virginia. May 2020.

SOUTHALL, B. L., BOWLES, A. E., ELLISON, W. T., FINNERAN, J. J., GENTRY, R. L., JR. GREENE, C. R., KASTAK, D., KETTEN, D. R., MILLER, J. H. et al. (2008). Marine mammal noise-exposure criteria: Initial scientific recommendations. *Bioacoustics* **17** 273–275.

SOUTHALL, B. L., FINNERAN, J. J., REICHMUTH, C., NACHTIGALL, P. E., KETTEN, D. R., BOWLES, A. E., ELLISON, W. T., NOWACEK, D. P. and TYACK, P. L. (2019). Marine mammal noise exposure criteria: Updated scientific recommendations for residual hearing effects. *Aquat. Mamm.* **45**.

SOUTHALL, B. L., NOWACEK, D. P., MILLER, P. J. and TYACK, P. L. (2016). Experimental field studies to measure behavioral responses of cetaceans to sonar. *Endanger. Species Res.* **31** 293–315.

STASINOPOULOS, D. M. and RIGBY, R. A. (2008). Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23** 1–46.

STIMPERT, A., DERUITER, S. L., SOUTHALL, B., MORETTI, D., FALCONE, E., GOLDBOGEN, J., FRIEDLAENDER, A., SCHORR, G. and CALAMBOKIDIS, J. (2014). Acoustic and foraging behavior of a Baird's beaked whale, Berardius bairdii, exposed to simulated sonar. *Sci. Rep.* **4** 1–8.

TYACK, P. L., ZIMMER, W. M., MORETTI, D., SOUTHALL, B. L., CLARIDGE, D. E., DURBAN, J. W., CLARK, C. W., D'AMICO, A., DIMARZIO, N. et al. (2011). Beaked whales respond to simulated and actual navy sonar. *PLoS ONE* **6** e17009.

WOOD, S. N. (2017). *Generalized Additive Models*: *An Introduction with R. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3726911