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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles—*Continued from front cover*

THE ANNALS OF APPLIED STATISTICS Vol. 18, No. 4, pp. 2679–3613 December 2024

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

President: Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104-6304, USA

President-Elect: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

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

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

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

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

IMS PUBLICATIONS

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

The Annals of Applied Statistics. *Editor-In-Chief*: Ji Zhu, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

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

The Annals of Applied Probability. *Editors*: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Qi-Man Shao, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, P.R. China

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

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

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 18, Number 4, December 2024. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2679–2704
https://doi.org/10.1214/24-AOAS1902

A COPULA MODEL FOR MARKED POINT PROCESS WITH A TERMINAL EVENT: AN APPLICATION IN DYNAMIC PREDICTION OF INSURANCE CLAIMS

BY LU YANG[1,a], PENG SHI[2,b] AND SHIMENG HUANG[2,c]

[1]*School of Statistics, University of Minnesota,* [a]*luyang@umn.edu*
[2]*Wisconsin School of Business, University of Wisconsin-Madison,* [b]*pshi@bus.wisc.edu,* [c]*shimeng.huang@wisc.edu*

Accurate prediction of an insurer's outstanding liabilities is crucial for maintaining the financial health of the insurance sector. We aim to develop a statistical model for insurers to dynamically forecast unpaid losses by leveraging the granular transaction data on individual claims. The liability cash flow from a single insurance claim is determined by an event process that describes the recurrences of payments, a payment process that generates a sequence of payment amounts, and a settlement process that terminates both the event and payment processes. More importantly, the three components are dependent on one another, which enables the dynamic prediction of an insurer's outstanding liability. We introduce a copula-based point process framework to model the recurrent events of payment transactions from an insurance claim, where the longitudinal payment amounts and the time-to-settlement outcome are formulated as the marks and the terminal event of the counting process, respectively. The dependencies among the three components are characterized using the method of pair copula constructions. We further develop a stage-wise strategy for parameter estimation and illustrate its desirable properties with numerical experiments.

In the application we consider a portfolio of property insurance claims for building and contents coverage obtained from a commercial property insurance provider, where we find intriguing dependence patterns among the three components. The superior dynamic prediction performance of the proposed joint model enhances the insurer's decision-making in claims reserving and risk financing operations.

REFERENCES

AALEN, O. O., BORGAN, Ø. and GJESSING, H. K. (2008). *Survival and Event History Analysis*: *A Process Point of View*. *Statistics for Biology and Health*. Springer, New York. MR2449233 https://doi.org/10.1007/978-0-387-68560-1

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884 https://doi.org/10.1016/j.insmatheco.2007.02.001

BEDFORD, T. and COOKE, R. M. (2002). Vines—a new graphical model for dependent random variables. *Ann. Statist.* **30** 1031–1068. MR1926167 https://doi.org/10.1214/aos/1031689016

BROWN, E. R., IBRAHIM, J. G. and DEGRUTTOLA, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61** 64–73. MR2129202 https://doi.org/10.1111/j.0006-341X.2005.030929.x

CAFFO, B. S., BOOTH, J. G. and DAVISON, A. C. (2002). Empirical supremum rejection sampling. *Biometrika* **89** 745–754. MR1946509 https://doi.org/10.1093/biomet/89.4.745

CHI, Y.-Y. and IBRAHIM, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62** 432–445. MR2227491 https://doi.org/10.1111/j.1541-0420.2005.00448.x

COOK, R. J. and LAWLESS, J. F. (2007). *The Statistical Analysis of Recurrent Events*. *Statistics for Biology and Health*. Springer, New York. MR3822124

DIAO, L., COOK, R. J. and LEE, K.-A. (2013). A copula model for marked point processes. *Lifetime Data Anal.* **19** 463–489. MR3119993 https://doi.org/10.1007/s10985-013-9259-3

Key words and phrases. Conditional copula, joint model, life cycle of insurance claims, longitudinal and survival outcomes.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 134–144.

DING, J. and WANG, J.-L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics* **64** 546–556. MR2432425 https://doi.org/10.1111/j.1541-0420.2007.00896.x

ELASHOFF, R., LI, N. et al. (2016). *Joint Models for Longitudinal and Time-to-Event Data*. CRC Press, Boca Raton.

FAUCETT, C. L. and THOMAS, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat. Med.* **15** 1663–1685.

GANJALI, M. and BAGHFALAKI, T. (2015). A copula approach to joint modeling of longitudinal measurements and survival times using Monte Carlo expectation-maximization with application to aids studies. *J. Biopharm. Statist.* **25** 1077–1099.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31** 1208–1211. MR0123385 https://doi.org/10.1214/aoms/1177705693

GRUTTOLA, V. D. and TU, X. M. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50** 1003–1014.

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.* **94** 401–419. MR2167922 https://doi.org/10.1016/j.jmva.2004.06.003

KIM, S., ZENG, D., CHAMBLESS, L. and LI, Y. (2012). Joint models of longitudinal data and recurrent events with informative terminal event. *Stat. Biosci.* **4** 262–281.

KRÓL, A., FERRER, L., PIGNON, J.-P., PROUST-LIMA, C., DUCREUX, M., BOUCHÉ, O., MICHIELS, S. and RONDEAU, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. *Biometrics* **72** 907–916. MR3545683 https://doi.org/10.1111/biom.12490

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LI, N., ELASHOFF, R. M., LI, G. and TSENG, C.-H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Stat. Med.* **31** 1707–1721. MR2947519 https://doi.org/10.1002/sim.4507

LIN, H., MCCULLOCH, C. E. and MAYNE, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Stat. Med.* **21** 2369–2382. https://doi.org/10.1002/sim.1179

LIU, L. and HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 65–81. MR2662234 https://doi.org/10.1111/j.1467-9876.2008.00641.x

LIU, L., HUANG, X. and O'QUIGLEY, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64** 950–958. MR2526647 https://doi.org/10.1111/j.1541-0420.2007.00954.x

MASAROTTO, G. and VARIN, C. (2012). Gaussian copula marginal regression. *Electron. J. Stat.* **6** 1517–1549. MR2988457 https://doi.org/10.1214/12-EJS721

MCDONALD, J. B. and XU, Y. J. (1995). A generalization of the beta distribution with applications. *J. Econometrics* **66** 133–152.

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, *Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu. Rev. Stat. Appl.* **6** 223–240. MR3939519 https://doi.org/10.1146/annurev-statistics-030718-105048

PARRY, M., DAWID, A. P. and LAURITZEN, S. (2012). Proper local scoring rules. *Ann. Statist.* **40** 561–592. MR3014317 https://doi.org/10.1214/12-AOS971

PENG, R. D. (2018). Advanced statistical computing. Work in Progress.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–342. MR0671971 https://doi.org/10.1093/biomet/69.2.331

PROUST-LIMA, C., DARTIGUES, J.-F. and JACQMIN-GADDA, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach. *Stat. Med.* **35** 382–398. MR3455508 https://doi.org/10.1002/sim.6731

RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*: *With Applications in R*. CRC press, Boca Raton.

RIZOPOULOS, D. and GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.* **30** 1366–1380. MR2828959 https://doi.org/10.1002/sim.4205

RIZOPOULOS, D., VERBEKE, G. and LESAFFRE, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 637–654. MR2749911 https://doi.org/10.1111/j.1467-9868.2008.00704.x

RIZOPOULOS, D., VERBEKE, G., LESAFFRE, E. and VANRENTERGHEM, Y. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics* **64** 611–619. MR2432435 https://doi.org/10.1111/j.1541-0420.2007.00894.x

RIZOPOULOS, D., VERBEKE, G. and MOLENBERGHS, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95** 63–74. MR2409715 https://doi.org/10.1093/biomet/asm087

SHI, P. (2014). Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science*, *Volume I*: *Predictive Modeling Techniques* (E. W. Edward, G. Meyers and R. A. Derrig, eds.) 236–259. Cambridge Univ. Press, Cambridge.

SHI, P. and YANG, L. (2018). Pair copula constructions for insurance experience rating. *J. Amer. Statist. Assoc.* **113** 122–133. MR3803444 https://doi.org/10.1080/01621459.2017.1330692

SURESH, K., TAYLOR, J. M. G. and TSODIKOV, A. (2021). A Gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker. *Biostatistics* **22** 504–521. MR4287165 https://doi.org/10.1093/biostatistics/kxz049

TANG, A.-M. and TANG, N.-S. (2015). Semiparametric Bayesian inference on skew-normal joint modeling of multivariate longitudinal and survival data. *Stat. Med.* **34** 824–843. MR3326393 https://doi.org/10.1002/sim.6373

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

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186 https://doi.org/10.2307/2533118

ZELLER, G. and SCHERER, M. (2022). A comprehensive model for cyber risk based on marked point processes and its application to insurance. *Eur. Actuar. J.* **12** 33–85. MR4443594 https://doi.org/10.1007/s13385-021-00290-1

ZHU, H., IBRAHIM, J. G., CHI, Y.-Y. and TANG, N. (2012). Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics* **68** 954–964. MR3055200 https://doi.org/10.1111/j.1541-0420.2012.01745.x

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2705–2729
https://doi.org/10.1214/24-AOAS1898

MODEL-ROBUST BAYESIAN DESIGN THROUGH GENERALISED ADDITIVE MODELS FOR MONITORING SUBMERGED SHOALS

BY DILISHIYA DE SILVA[1,a], REBECCA FISHER[2,d], BEN RADFORD[2,e], HELEN THOMPSON[1,b] AND JAMES MCGREE[1,c]

[1]*School of Mathematical Sciences, Queensland University of Technology,* [a]*dilishiya.desilva@qut.edu.au,* [b]*helen.thompson@qut.edu.au,* [c]*james.mcgree@qut.edu.au*

[2]*Australian Institute of Marine Science, UWA Oceans Institute,* [d]*r.fisher@aims.gov.au,* [e]*b.radford@aims.gov.au*

Optimal sampling strategies are critical for surveys of deeper coral reef and shoal systems due to the significant cost of accessing and field sampling these remote and poorly understood ecosystems. Additionally, well-established standard diver-based sampling techniques used in shallow reef systems are not feasible at greater depths. In this study, we develop a Bayesian design strategy to optimise sampling for a shoal deep reef system using three years of pilot data. Bayesian designs are typically found by maximising the expectation of a utility function with respect to the joint distribution of the parameters and the response conditional on an assumed statistical model. Unfortunately, specifying such a model a priori is difficult, as knowledge of the data-generating process is typically incomplete. To overcome this, our approach focuses on finding Bayesian designs that are robust to unknown model uncertainty. We achieve this by couching the specified model within a generalised additive modelling framework and formulating prior information that allows the additive component to capture discrepancies between what is assumed and the underlying data-generating process. The motivation for this is to enable Bayesian designs to be found under epistemic model uncertainty; a highly desirable property of Bayesian designs. Initially, we demonstrate our approach with an exemplar design problem, deriving a theoretical result to explore the properties of optimal designs. We then apply this approach to design future monitoring of submerged shoals off the north-west coast of Australia to improve current monitoring practices.

REFERENCES

ABDUL WAHAB, M. A., RADFORD, B., CAPPO, M., COLQUHOUN, J., STOWAR, M., DEPCZYNSKI, M., MILLER, K. and HEYWARD, A. (2018). Biodiversity and spatial patterns of benthic habitat and associated demersal fish communities at two tropical submerged reef ecosystems. *Coral Reefs* **37** 327–343. https://doi.org/10.1007/s00338-017-1655-9

ALDRIN, M. and HAFF, I. H. (2005). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmos. Environ.* **39** 2145–2155. https://doi.org/10.1016/j.atmosenv.2004.12.020

ANYOSA, S., EIDSVIK, J. and PIZARRO, O. (2023). Adaptive spatial designs minimizing the integrated Bernoulli variance in spatial logistic regression models - with an application to benthic habitat mapping. *Comput. Statist. Data Anal.* **179** Paper No. 107643. MR4497906 https://doi.org/10.1016/j.csda.2022.107643

ATKINSON, A. C., DONEV, A. N. and TOBIAS, R. D. (2007). *Optimum Experimental Designs, with SAS. Oxford Statistical Science Series* **34**. Oxford Univ. Press, Oxford. MR2323647

BARNES, R. S. K. and HUGHES, R. N. (1999). *An Introduction to Marine Ecology*. Wiley, New York.

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Stat.* **37** 51–58. MR0189176 https://doi.org/10.1214/aoms/1177699597

BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory. Wiley Series in Probability and Statistics*. Wiley, Chichester, England. MR1274699

BIEDERMANN, S., DETTE, H. and WOODS, D. C. (2009). Optimal designs for multivariable spline models. (S3RI Methodology Working Papers; No. M09/16), Univ. Southampton, Southampton Statistical Sciences Research Institute.

Key words and phrases. Bayesian model selection, Laplace approximation, model uncertainty, O'Sullivan penalised splines, robust design, sampling design, transect design.

BIEDERMANN, S., DETTE, H. and WOODS, D. C. (2011). Optimal design for additive partially nonlinear models. *Biometrika* **98** 449–458. MR2806440

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

BORNKAMP, B., BRETZ, F., DETTE, H. and PINHEIRO, J. (2011). Response-adaptive dose-finding under model uncertainty. *Ann. Appl. Stat.* **5** 1611–1631. MR2849788 https://doi.org/10.1214/10-AOAS445

BOS, C. S. (2002). A comparison of marginal likelihood computation methods. In *COMPSTAT* 2002 (*Berlin*) 111–116. Physica, Heidelberg. MR1973480

BRIDGE, T. C., HUGHES, T. P., GUINOTTE, J. M. and BONGAERTS, P. (2013). Call to protect all coral reefs. *Nat. Clim. Change* **3** 528–530. https://doi.org/10.1038/nclimate1879

CHANG, Y.-J. and NOTZ, W. I. (1996). Model robust designs. In *Design and Analysis of Experiments. Handbook of Statist.* **13** 1055–1098. North-Holland, Amsterdam. MR1492590

COOK, R. D. and NACHTSHEIM, C. J. (1982). Model robust, linear-optimal designs. *Technometrics* **24** 49–54. MR0653111 https://doi.org/10.2307/1267577

CURE, K., CURREY-RANDALL, L., GALAIDUK, R., RADFORD, B., WAKEFORD, M. and HEYWARD, A. (2021). Depth gradients in abundance and functional roles suggest limited depth refuges for herbivorous fishes. *Coral Reefs* **40** 365–379. https://doi.org/10.1007/s00338-021-02060-7

DE SILVA, D., FISHER, R., RADFORD, B., THOMPSON, H. and MCGREE, J. (2024). Supplement to "Model-robust Bayesian design through generalised additive models for monitoring submerged shoals." https://doi.org/10.1214/24-AOAS1898SUPPA, https://doi.org/10.1214/24-AOAS1898SUPPB, https://doi.org/10.1214/24-AOAS1898SUPPC

DEGROOT, M. H. (1962). Uncertainty, information, and sequential experiments. *Ann. Math. Stat.* **33** 404–419. MR0139242 https://doi.org/10.1214/aoms/1177704567

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

DETTE, H., MELAS, V. B. and PEPELYSHEV, A. (2008). Optimal designs for free knot least squares splines. *Statist. Sinica* **18** 1047–1062. MR2440403

DETTE, H., MELAS, V. B. and PEPELYSHEV, A. (2011). Optimal design for smoothing splines. *Ann. Inst. Statist. Math.* **63** 981–1003. MR2822964 https://doi.org/10.1007/s10463-009-0265-x

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

DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2014). A sequential Monte Carlo algorithm to incorporate model uncertainty in Bayesian sequential design. *J. Comput. Graph. Statist.* **23** 3–24. MR3173758 https://doi.org/10.1080/10618600.2012.730083

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

FALKOWSKI, P. G., DUBINSKY, Z., MUSCATINE, L. and PORTER, J. W. (1984). Light and the bioenergetics of a symbiotic coral. *BioScience* **34** 705–709. https://doi.org/10.2307/1309663

FOSSUM, T. O., TRAVELLETTI, C., EIDSVIK, J., GINSBOURGER, D. and RAJAN, K. (2021). Learning excursion sets of vector-valued Gaussian random fields for autonomous ocean sampling. *Ann. Appl. Stat.* **15** 597–618. MR4298973 https://doi.org/10.1214/21-aoas1451

FOSTER, S. D. (2021). MBHdesign: An R-package for efficient spatial survey designs. *Methods Ecol. Evol.* **12** 415–420. https://doi.org/10.1111/2041-210X.13535

FOSTER, S. D., HOSACK, G. R., MONK, J., LAWRENCE, E., BARRETT, N. S., WILLIAMS, A. and PRZESLAWSKI, R. (2020). Spatially balanced designs for transect-based surveys. *Methods Ecol. Evol.* **11** 95–105. https://doi.org/10.1111/2041-210X.13321

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492

HASTIE, T. and TIBSHIRANI, R. (1987). Generalized additive models: some applications. *J. Amer. Statist. Assoc.* **82** 371–386. MR0858512 https://doi.org/10.2307/2289439

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

HEILIGERS, B. (1998). E-optimal designs for polynomial spline regression. *J. Statist. Plann. Inference* **75** 159–172. MR1671698

HEILIGERS, B. (1999). Experimental design for polynomial spline regression. *Tatra Mt. Math. Publ.* **17** 157–165.

HEYWARD, A., CASE, M., CAPPO, M., COLQUHOUN, J., CURRY, L., FISHER, R., RADFORD, B., STOWAR, M., WAKEFORD, M. et al. (2017). The Barracouta, Goeree and Vulcan, Shoals Survey 2016. Report, Townsville: Australian Institute of Marine Science.

HEYWARD, A., JONES, R., MEEUWIG, J., BURNS, K., RADFORD, B., COLQUHOUN, J., CAPPO, M., CASE, M., O'LEARY, R. et al. (2012). Montara: 2011 Offshore banks assessment survey. Report for PTTEP Australasia (Ashmore Cartier) Pty. Ltd. Australian Institute of Marine Science, Townsville, Australia.

HEYWARD, A. and RADFORD, B. (2019). Northwest Australia. In *Mesophotic Coral Ecosystems* 337–349. Springer, Berlin.

HUGHES, T. P., ANDERSON, K. D., CONNOLLY, S. R., HERON, S. F., KERRY, J. T., LOUGH, J. M., BAIRD, A. H., BAUM, J. K., BERUMEN, M. L. et al. (2018). Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359** 80–83. https://doi.org/10.1126/science.aan8048

KAHNG, S., GARCIA-SAIS, J., SPALDING, H., BROKOVICH, E., WAGNER, D., WEIL, E., HINDERSTEIN, L. and TOONEN, R. (2010). Community ecology of mesophotic coral reef ecosystems. *Coral Reefs* **29** 255–275. https://doi.org/10.1007/s00338-010-0593-6

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

KRISTOFFERSEN, P. and SMUCKER, B. J. (2020). Model-robust design of mixture experiments. *Qual. Eng.* **32** 663–675. https://doi.org/10.1080/08982112.2020.1722831

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

LÄUTER, E. (1974). Experimental design in a class of models. *Math. Operationsforsch. Statist.* **5** 379–398. MR0440812 https://doi.org/10.1080/02331887408801175

LÄUTER, E. (1976). Optimal multipurpose designs for regression models. *Math. Operationsforsch. Statist.* **7** 51–68. MR0413385 https://doi.org/10.1080/02331887608801276

LAVERICK, J. H., TAMIR, R., EYAL, G. and LOYA, Y. (2020). A generalized light-driven model of community transitions along coral reef depth gradients. *Glob. Ecol. Biogeogr.* **29** 1554–1564.

LIN, X. and ZHANG, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 381–400. MR1680318 https://doi.org/10.1111/1467-9868.00183

LINDLEY, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27** 986–1005. MR0083936 https://doi.org/10.1214/aoms/1177728069

MATÉRN, B. (1960). *Spatial Variation* : *Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. *Stockholm. Statens Skogsforskningsinstitut. Meddelanden*. University of Sweden.

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

MEYER, R. K. and NACHTSHEIM, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics* **37** 60–69. MR1322047 https://doi.org/10.2307/1269153

MOORE, C., CAPPO, M., RADFORD, B. and HEYWARD, A. (2017). Submerged oceanic shoals of North western Australia are a major reservoir of marine biodiversity. *Coral Reefs* **36** 719–734.

MOORE, C., DRAZEN, J. C., RADFORD, B. T., KELLEY, C. and NEWMAN, S. J. (2016). Improving essential fish habitat designation to support sustainable ecosystem-based fisheries management. *Mar. Policy* **69** 32–41. https://doi.org/10.1016/j.marpol.2016.03.021

MURASE, H., NAGASHIMA, H., YONEZAKI, S., MATSUKURA, R. and KITAKADO, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: A case study in Sendai Bay, Japan. *ICES J. Mar. Sci.* **66** 1417–1424. https://doi.org/10.1093/icesjms/fsp105

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–518. MR0874480 https://doi.org/10.1214/ss/1177013525

OBURA, D. O., AEBY, G., AMORNTHAMMARONG, N., APPELTANS, W., BAX, N., BISHOP, J., BRAINARD, R. E., CHAN, S., FLETCHER, P. et al. (2019). Coral reef monitoring, reef assessment technologies, and ecosystem-based management. *Front. Mar. Sci.* **6** 580. https://doi.org/10.3389/fmars.2019.00580

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. MR2370838 https://doi.org/10.1198/016214506000001491

OVERSTALL, A. M., MCGREE, J. M. and DROVANDI, C. C. (2018). An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Stat. Comput.* **28** 343–358. MR3747567 https://doi.org/10.1007/s11222-017-9734-x

OVERSTALL, A. M. and WOODS, D. C. (2017). Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59** 458–470. MR3740963 https://doi.org/10.1080/00401706.2016.1251495

PEARCE, J. L., BERINGER, J., NICHOLLS, N., HYNDMAN, R. J. and TAPPER, N. J. (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmos. Environ.* **45** 1328–1336. https://doi.org/10.1016/j.atmosenv.2010.11.051

PEDERSEN, E. J., MILLER, D. L., SIMPSON, G. L. and ROSS, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ* **7** e6876. https://doi.org/10.7717/peerj.6876

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

RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2017). Bayesian computing with INLA: A review. *Annu. Rev. Stat. Appl.* **4** 395–421.

RYAN, E. G., DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.* **84** 128–154.

SACKS, J. and YLVISAKER, D. (1984). Some model robust designs in regression. *Ann. Statist.* **12** 1324–1348. MR0760692 https://doi.org/10.1214/aos/1176346795

SENARATHNE, S. G. J., OVERSTALL, A. M. and MCGREE, J. M. (2020). Bayesian adaptive N-of-1 trials for estimating population and individual treatment effects. *Stat. Med.* **39** 4499–4518. MR4175093

SWARTZMAN, G. (1997). Analysis of the summer distribution of fish schools in the Pacific eastern boundary current. *ICES Journal of Marine Science* **54** 105–116.

TAYLOR, J. C. and RAND, P. S. (2003). Spatial overlap and distribution of anchovies (anchoa spp.) and copepods in a shallow stratified estuary. *Aquatic Living Resour.* **16** 191–196.

WAND, M. P. and ORMEROD, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Aust. N. Z. J. Stat.* **50** 179–198. MR2431193 https://doi.org/10.1111/j.1467-842X.2008.00507.x

WANG, J., VERBYLA, A. P., JIANG, B., ZWART, A. B., ONG, C. S., SIRAULT, X. R. R. and VERBYLA, K. L. (2020). Optimal design for adaptive smoothing splines. *J. Statist. Plann. Inference* **206** 263–277. MR4036707 https://doi.org/10.1016/j.jspi.2019.10.002

WELCH, W. J. (1983). A mean squared error criterion for the design of experiments. *Biometrika* **70** 205–213. MR0742990 https://doi.org/10.1093/biomet/70.1.205

WINES, S. L., YOUNG, M. A., ZAVALAS, R., LOGAN, J. M., TINKLER, P. and IERODIACONOU, D. (2020). Accounting for spatial scale and temporal variation in fish-habitat analyses using baited remote underwater video stations (BRUVS). *Mar. Ecol. Prog. Ser.* **640** 171–187. https://doi.org/10.3354/meps13292

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

WOOD, S. N. (2016). Just another Gibbs additive modeler: Interfacing JAGS and mgcv. *J. Stat. Softw.* **75** 1–15. https://doi.org/10.18637/jss.v075.i07

WOOD, S. N. (2017). *Generalized Additive Models*: *An Introduction with R. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3726911

WOODS, D. C., LEWIS, S. M. and DEWYNNE, J. N. (2003). Designing experiments for multi-variable B-spline models. *Sankhyā* **65** 660–677. MR2060614

XU, X. and SINHA, S. K. (2021). Robust designs for generalized linear mixed models with possible model misspecification. *J. Statist. Plann. Inference* **210** 20–41. MR4101484

YATES, K. L., BOUCHET, P. J., CALEY, M. J., MENGERSEN, K., RANDIN, C. F., PARNELL, S., FIELDING, A. H., BAMFORD, A. J., BAN, S. et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* **33** 790–802. https://doi.org/10.1016/j.tree.2018.08.001

YEE, T. W. and MITCHELL, N. D. (1991). Generalized additive models in plant ecology. *J. Veg. Sci.* **2** 587–602. https://doi.org/10.2307/3236170

ZHOU, X., JOSEPH, L., WOLFSON, D. B. and BÉLISLE, P. (2003). A Bayesian *A*-optimal and model robust design criterion. *Biometrics* **59** 1082–1088. MR2025133 https://doi.org/10.1111/j.0006-341X.2003.00124.x

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2730–2752
https://doi.org/10.1214/24-AOAS1903

REGULARIZED SCALAR-ON-FUNCTION REGRESSION ANALYSIS TO ASSESS FUNCTIONAL ASSOCIATION OF CRITICAL PHYSICAL ACTIVITY WINDOW WITH BIOLOGICAL AGE

BY MARGARET BANKER[a], LEYAO ZHANG[b] AND PETER X. K. SONG[c]

Department of Biostatistics, University of Michigan, [a]*mbanker@umich.edu,* [b]*leyaozh@umich.edu,* [c]*pxsong@umich.edu*

Accelerometry data enables scientists to extract personal digital features useful in precision health decision making. Existing analytic methods often begin with discretizing physical activity (PA) counts into activity categories via fixed cutoffs; however, the cutoffs are validated under restricted settings and cannot be generalized across studies. Here we develop a data-driven approach to overcome this bottleneck in the analysis of PA data in which we holistically summarize an individual's PA profile using occupation-time curves that describe the percentage of time spent at or above a continuum of activity levels. The resulting functional curve is informative to capture time-course individual variability of PA. We investigate functional analytics under an L_0 regularization approach, which handles highly correlated micro-activity windows that serve as predictors in a scalar-on-function regression model. We develop a new one-step method that simultaneously conducts fusion via change-point detection and parameter estimation through a new L_0 constraint formulation, which is evaluated via simulation experiments and a data analysis assessing the influence of PA on biological aging.

REFERENCES

BANDE, M. F., FUENTE, M. O. D. L., GALEANO, P., NIETO, A. and GARCIA-PORTUGUES, E. (2022). Fda.usc: Functional data analysis and utilities for statistical computing.

BANKER, M., ZHANG, L. and SONG, P. X. (2024). Supplement to "Regularized scalar-on-function regression analysis to assess functional association of critical physical activity window with biological age." https://doi.org/10.1214/24-AOAS1903SUPP

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

BERTSIMAS, D., PAUPHILET, J. and VAN PARYS, B. (2020). Sparse regression: Scalable algorithms and empirical performance. *Statist. Sci.* **35** 555–578. MR4175381 https://doi.org/10.1214/19-STS701

BERTSIMAS, D. and SHIODA, R. (2009). Algorithm for cardinality-constrained quadratic optimizaiton. *Comput. Optim. Appl.* **43** 1–22. MR2501042 https://doi.org/10.1007/s10589-007-9126-9

BOGACHEV, L. and RATANOV, N. (2011). Occupation time distributions for the telegraph process. *Stochastic Process. Appl.* **121** 1816–1844. MR2811025 https://doi.org/10.1016/j.spa.2011.03.016

CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688 https://doi.org/10.1214/08-AOS653

CANDÈS, E. J., WAKIN, M. B. and BOYD, S. P. (2008). Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Appl.* **14** 877–905. MR2461611 https://doi.org/10.1007/s00041-008-9045-x

CHANDLER, J. L., BRAZENDALE, K., BEETS, M. W. and MEALING, B. A. (2016). Classification of physical activity intensities using a wrist-worn accelerometer in 8-12-year-old children. *Int. J. Pediatr. Obes.* **11** 120–127. https://doi.org/10.1111/ijpo.12033

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

CHEN, K. Y. and BASSETT, D. R. (2005). The technology of accelerometry-based activity monitors: Current and future. *Med. Sci. Sports Exerc.* **37** S490–500.

CROUTER, S. E., FLYNN, J. I. and BASSETT, D. R. (2015). Estimating physical activity in youth using a wrist accelerometer. *Med. Sci. Sports Exerc.* **47** 944–951. https://doi.org/10.1249/MSS.0000000000000502

Key words and phrases. Functional data analysis, fusion regularization, occupation time curve, scalar-on-function regression, wearable device.

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

FREEDSON, P., POBER, D. and JANZ, K. F. (2005). Calibration of accelerometer output for children. *Med. Sci. Sports Exerc.* **37** S523–S530. https://doi.org/10.1249/01.mss.0000185658.28284.ba

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 453–469. MR2914521 https://doi.org/10.1111/j.1467-9876.2011.01031.x

GOLDSMITH, J., SCHEIPL, F., HUANG, L., WROBEL, J., DI, C., GELLAR, J., HAREZLAK, J., MCLEAN, M. W., SWIHART, B. et al. (2024). refund: Regression with functional data.

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

HEINZL, F. and TUTZ, G. (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biom. J.* **56** 44–68. MR3152702 https://doi.org/10.1002/bimj.201200111

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

HORVATH, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* **14** 3156.

HORVATH, S., OSHIMA, J., MARTIN, G. M., LU, A. T., QUACH, A., COHEN, H., FELTON, S., MATSUYAMA, M., LOWE, D. et al. (2018). Epigenetic clock for skin and blood cells applied to Hutchinson Gilford progeria syndrome and. *Aging* **10** 1758–1775. https://doi.org/10.18632/aging.101508

HUANG, R.-C., LILLYCROP, K. A., BEILIN, L. J., GODFREY, K. M., ANDERSON, D., MORI, T. A., RAUSCHERT, S., CRAIG, J. M., ODDY, W. H. et al. (2019). Epigenetic age acceleration in adolescence associates with BMI, inflammation, and risk score for middle age cardiovascular disease. *J. Clin. Endocrinol. Metab.* **104** 3012–3024.

KANKAANPÄÄ, A., TOLVANEN, A., HEIKKINEN, A., KAPRIO, J., OLLIKAINEN, M. and SILLANPÄÄ, E. (2022). The role of adolescent lifestyle habits in biological aging: A prospective twin study. *eLife* **11**. https://doi.org/10.7554/eLife.80729

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

MARIONI, R. E., SHAH, S., MCRAE, A. F., RITCHIE, S. J., MUNIZ-TERRERA, G., HARRIS, S. E., GIBSON, J., REDMOND, P., COX, S. R. et al. (2015). The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int. J. Epidemiol.* **44** 1388–1396.

MCEWEN, L. M., O'DONNELL, K. J., MCGILL, M. G., EDGAR, R. D., JONES, M. J., MACISAAC, J. L., LIN, D. T. S., RAMADORI, K., MORIN, A. et al. (2020). The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc. Natl. Acad. Sci. USA* **117** 23329–23335. https://doi.org/10.1073/pnas.1820843116

MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. *Monographs on Statistics and Applied Probability* **95**. CRC Press, Boca Raton, FL. MR2001193 https://doi.org/10.1201/9781420035933

NATARAJAN, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.* **24** 227–234. MR1320206 https://doi.org/10.1137/S0097539792240406

NWANAJI-ENWEREM, J. C., LAAN, L. V. D., KOGUT, K., ESKENAZI, B., HOLLAND, N., DEARDORFF, J. and CARDENAS, A. (2021). Maternal adverse childhood experiences before pregnancy are associated with epigenetic aging changes in their children. *Aging* **13** 25653–25669.

PERNG, W., TAMAYO-ORTIZ, M., TANG, L., SÁNCHEZ, B. N., CANTORAL, A., MEEKER, J. D., DOLINOY, D. C., ROBERTS, E. F., MARTINEZ-MIER, E. A. et al. (2019). Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT) project. *BMJ Open* **9** e030427. https://doi.org/10.1136/bmjopen-2019-030427

QUACH, A., LEVINE, M. E., TANAKA, T., LU, A. T., CHEN, B. H., FERRUCCI, L., RITZ, B., BANDINELLI, S., NEUHOUSER, M. L. et al. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging* **9** 419–437.

RABINER, L. and JUANG, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Mag.* **3** 4–16.

RAMSAY, J. (2005). *Functional Data Analysis*. *Encyclopedia of Statistics in Behavioral Science*. Wiley, New York.

RAMSAY, J. O. (2004). *Functional Data Analysis*. *Encyclopedia of Statistical Sciences*. Wiley, New York.

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

REISS, P. T., GOLDSMITH, J., SHANG, H. L. and OGDEN, R. T. (2017). Methods for scalar-on-function regression. *Int. Stat. Rev.* **85** 228–249. MR3686566 https://doi.org/10.1111/insr.12163

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

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

VRIEZE, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* **17** 228–243. https://doi.org/10.1037/a0027127

WANG, W., WU, S., ZHU, Z., ZHOU, L. and SONG, P. X.-K. (2024). Supervised homogeneity fusion: A combinatorial approach. *Ann. Statist.* **52** 285–310. MR4718416 https://doi.org/10.1214/23-aos2347

WIKLUND, P., KARHUNEN, V., RICHMOND, R. C., PARMAR, P., RODRIGUEZ, A., DE SILVA, M., WIELSCHER, M., REZWAN, F. I., RICHARDSON, T. G. et al. (2019). DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. *Clin. Epigenetics* **11** 97.

WU, X., CHEN, W., LIN, F., HUANG, Q., ZHONG, J., GAO, H., SONG, Y. and LIANG, H. (2019). DNA methylation profile is a quantitative measure of biological aging in children. *Aging* **11** 10031–10051.

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

ZHU, Y., SHEN, X. and PAN, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *J. Amer. Statist. Assoc.* **108** 713–725. MR3174654 https://doi.org/10.1080/01621459.2013.770704

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

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2753–2774
https://doi.org/10.1214/24-AOAS1904

CAUSAL HEALTH IMPACTS OF POWER PLANT EMISSION CONTROLS UNDER MODELED AND UNCERTAIN PHYSICAL PROCESS INTERFERENCE

BY NATHAN B. WIKLE[1,a] AND CORWIN M. ZIGLER[2,b]

[1]*Department of Statistics and Actuarial Science, University of Iowa,* [a]*nathan-wikle@uiowa.edu*
[2]*Department of Biostatistics, Brown University School of Public Health,* [b]*corwin_zigler@brown.edu*

Causal inference with spatial environmental data is often challenging due to the presence of interference: outcomes for observational units depend on some combination of local and nonlocal treatment. This is especially relevant when estimating the effect of power plant emissions controls on population health, as pollution exposure is dictated by: (i) the location of point-source emissions as well as (ii) the transport of pollutants across space via dynamic physical-chemical processes. In this work we estimate the effectiveness of air quality interventions at coal-fired power plants in reducing two adverse health outcomes in Texas in 2016: pediatric asthma ED visits and Medicare all-cause mortality. We develop methods for causal inference with interference when the underlying network structure is not known with certainty and instead must be estimated from ancillary data. Notably, uncertainty in the interference structure is propagated to the resulting causal effect estimates. We offer a Bayesian, spatial mechanistic model for the interference mapping, which we combine with a flexible nonparametric outcome model to marginalize estimates of causal effects over uncertainty in the structure of interference. Our analysis finds some evidence that emissions controls at upwind power plants reduce asthma ED visits and all-cause mortality; however, accounting for uncertainty in the interference renders the results largely inconclusive.

REFERENCES

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

AUSTIN, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28** 3083–3107. MR2750408 https://doi.org/10.1002/sim.3697

AUSTIN, P. C. (2019). Assessing covariate balance when using the generalized propensity score with quantitative or continuous exposures. *Stat. Methods Med. Res.* **28** 1365–1377. MR3941081 https://doi.org/10.1177/0962280218756159

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

BLANGIARDO, M., FINAZZI, F. and CAMELETTI, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions: Environmental exposure and health. *Spat. Spatio-Tempor. Epidemiol.* **18** 1–12.

BUONOCORE, J. J., DONG, X., SPENGLER, J. D., FU, J. S. and LEVY, J. I. (2014). Using the Community Multiscale Air Quality (CMAQ) model to estimate public health impacts of PM2.5 from individual power plants. *Environ. Int.* **68** 200–208.

CASEY, J. A., SU, J. G., HENNEMAN, L. R. F., ZIGLER, C., NEOPHYTOU, A. M., CATALANO, R., GONDALIA, R., CHEN, Y.-T., KAYE, L. et al. (2020). Improved asthma outcomes observed in the vicinity of coal power plant retirement, retrofit and conversion to natural gas. *Nat. Energy* **5** 398–408.

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

CLIFF, A. D. and ORD, J. K. (1981). *Spatial Processes: Models & Applications*. Pion Ltd., London. MR0632256

Key words and phrases. Causal inference, interference, air pollution, mechanistic models, BART.

COMESS, S., CHANG, H. H. and WARREN, J. L. (2024). A Bayesian framework for incorporating exposure uncertainty into health analyses with application to air pollution and stillbirth. *Biostatistics* **25** 20–39. MR4678529 https://doi.org/10.1093/biostatistics/kxac034

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

CULLIS, C. F. and HIRSCHLER, M. M. (1980). Atmospheric sulphur: Natural and man-made sources. *Atmos*. *Environ*. **14** 1263–1278.

DÍAZ MUÑOZ, I. and VAN DER LAAN, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics* **68** 541–549. MR2959621 https://doi.org/10.1111/j.1541-0420.2011.01685.x

DIGGLE, P. and ELLIOTT, P. (1995). Disease risk near point sources: Statistical issues for analyses using individual or spatially aggregated data. *J. Epidemiol*. *Community Health* **49** S20–S27.

DOMINICI, F., GREENSTONE, M. and SUNSTEIN, C. R. (2014). Science and regulation. Particulate matter matters. *Science* **344** 257–259. https://doi.org/10.1126/science.1247348

DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist*. *Sci*. **34** 43–68. MR3938963 https://doi.org/10.1214/18-STS667

DOUDCHENKO, N., ZHANG, M., DRYNKIN, E., AIROLDI, E., MIRROKNI, V. and POUGET-ABADIE, J. (2020). Causal inference with bipartite designs.

FOLEY, K. M., NAPELENOK, S. L., JANG, C., PHILLIPS, S., HUBBELL, B. J. and FULCHER, C. M. (2014). Two reduced form air quality modeling techniques for rapidly calculating pollutant mitigation potential across many sources, locations and precursor emission types. *Atmos*. *Environ*. **98** 283–289.

FONG, C., HAZLETT, C. and IMAI, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann*. *Appl*. *Stat*. **12** 156–177. MR3773389 https://doi.org/10.1214/17-AOAS1101

FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *J*. *Amer*. *Statist*. *Assoc*. **116** 901–918. MR4270033 https://doi.org/10.1080/01621459.2020.1768100

FORASTIERE, L., MEALLI, F. and VANDERWEELE, T. J. (2016). Identification and estimation of causal mechanisms in clustered encouragement designs: Disentangling bed nets using Bayesian principal stratification. *J*. *Amer*. *Statist*. *Assoc*. **111** 510–525. MR3538683 https://doi.org/10.1080/01621459.2015.1125788

FORASTIERE, L., MEALLI, F., WU, A. and AIROLDI, E. M. (2022). Estimating causal effects under network interference with Bayesian generalized propensity scores. *J*. *Mach*. *Learn*. *Res*. **23** Paper No. 289. MR4577728

GARCIA, E., RICE, M. B. and GOLD, D. R. (2021). Air pollution and lung function in children. *J*. *Allergy Clin*. *Immunol*. **148** 1–14. https://doi.org/10.1016/j.jaci.2021.05.006

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

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal*. **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

HENNEMAN, L., CHOIRAT, C., DEDOUSSI, I., DOMINICI, F., ROBERTS, J. and ZIGLER, C. (2023). Mortality risk from United States coal electricity generation. *Science* **382** 941–946.

HENNEMAN, L. R. F., CHOIRAT, C., IVEY, C., CUMMISKEY, K. and ZIGLER, C. M. (2019). Characterizing population exposure to coal emissions sources in the United States using the HyADS model. *Atmos*. *Environ*. **203** 271–280.

HENNEMAN, L. R. F., CHOIRAT, C. and ZIGLER, C. M. (2019). Accountability assessment of health improvements in the United States associated with reduced coal emissions between 2005 and 2012. *Epidemiology* **30** 477–485.

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

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J*. *Amer*. *Statist*. *Assoc*. **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

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

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **76** 243–263. MR3153941 https://doi.org/10.1111/rssb.12027

JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules.

KALNAY, E., KANAMITSU, M. et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77** 437–470.
KARWA, V., AIROLDI, E. M. (2018). A systematic investigation of classical causal inference strategies under mis-specification due to network interference.
LI, X., HAN, J., LIU, Y., DOU, Z. and AN ZHANG, T. (2022). Summary of research progress on industrial flue gas desulfurization technology. *Separation and Purification Technology* **281** 119849.
LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564 https://doi.org/10.1080/01621459.2013.844698
MCCLURE, M., GIBSON, R., CHIU, K.-K. and RANGANATH, R. (2017). Identifying potentially induced seismicity and assessing statistical significance in Oklahoma and California. *J. Geophys. Res., Solid Earth* **122** 2153–2172.
MURRAY, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *J. Amer. Statist. Assoc.* **116** 756–769. MR4270022 https://doi.org/10.1080/01621459.2020.1813587
OHNISHI, Y., KARMAKAR, B. and SABBAGHI, A. (2023). Degree of interference: A general framework for causal inference under interference.
ORELLANO, P., REYNOSO, J. and QUARANTA, N. (2021). Short-term exposure to sulphur dioxide (SO. *Environ. Int.* **150** 106434. https://doi.org/10.1016/j.envint.2021.106434
PEREZ-HEYDRICH, C., HUDGENS, M. G., HALLORAN, M. E., CLEMENS, J. D., ALI, M. and EMCH, M. E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics* **70** 734–744. MR3261790 https://doi.org/10.1111/biom.12184
PLUMMER, M. (2015). Cuts in Bayesian graphical models. *Stat. Comput.* **25** 37–43. MR3304902 https://doi.org/10.1007/s11222-014-9503-z
POLLMANN, M. (2023). Causal inference for spatial treatments.
POPE, C. A., EZZATI, M. and DOCKERY, D. W. (2009). Fine-particulate air pollution and life expectancy in the United States. *N. Engl. J. Med.* **360** 376–386. https://doi.org/10.1056/NEJMsa0805646
QU, Z., XIONG, R., LIU, J. and IMBENS, G. (2022). Efficient treatment effect estimation in observational studies under heterogeneous partial interference.
RACKAUCKAS, C., MA, Y., MARTENSEN, J., WARNER, C., ZUBOV, K., SUPEKAR, R., SKINNER, D., RAMADHAN, A. and EDELMAN, A. (2021). Universal differential equations for scientific machine learning.
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
RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696
SÄVJE, F. (2024). Causal inference with misspecified exposure mappings: Separating definitions and assumptions. *Biometrika* **111** 1–15. MR4704553 https://doi.org/10.1093/biomet/asad019
SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636
TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779
THCIC (2022). Texas Hospital Emergency Department Research Data File. Available at https://www.dshs.texas.gov/thcic/Texas-Hospital-Emergency-Department-Research-Data-File-(ED-RDF)/.
UDS MAPPER (2022). UDS Mapper: ZIP Code to ZCTA Crosswalk. Available at https://udsmapper.org/zip-code-to-zcta-crosswalk/.
UHLENBECK, G. E. and ORNSTEIN, L. S. (1930). On the theory of the Brownian motion. *Phys. Rev.* **36** 823–841.
US EPA (2003). Latest Findings on National Air Quality: 2002 Status and Trends. Available at https://www.epa.gov/air-trends/historical-air-quality-trends-reports.
US EPA (2013). America's Children and the Environment. Third Edition. Available at https://www.epa.gov/americaschildrenenvironment/americas-children-and-environment-third-edition.
US EPA (2016). Air Markets Program Data. Available at https://ampd.epa.gov/ampd/.
VAN DONKELAAR, A., MARTIN, R. V., LI, C. and BURNETT, R. T. (2019). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environ. Sci. Technol.* **53** 2595–2611.
VAN DER LAAN, M. J. (2014). Causal inference for a population of causally connected units. *J. Causal Inference* **2** 13–74. MR4289412 https://doi.org/10.1515/jci-2013-0002
WANG, Y., SAMII, C., CHANG, H. and ARONOW, P. M. (2023). Design-based inference for spatial experiments under unknown interference.

WEHNER, M. (2023). Connecting extreme weather events to climate change. *Phys. Today* **76** 40–46.
WIKLE, C. K. and HOOTEN, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *TEST* **19** 417–451. MR2745992 https://doi.org/10.1007/s11749-010-0209-z
WIKLE, N. B., HANKS, E. M., HENNEMAN, L. R. F. and ZIGLER, C. M. (2022). A mechanistic model of annual sulfate concentrations in the United States. *J. Amer. Statist. Assoc.* **117** 1082–1093. MR4480692 https://doi.org/10.1080/01621459.2022.2027774
WIKLE, N. B. and ZIGLER, C. M. (2024). Supplement to "Causal health impacts of power plant emission controls under modeled and uncertain physical process interference." https://doi.org/10.1214/24-AOAS1904SUPP
XIA, Y., ZHANG, M., TSANG, D. C. W., GENG, N., LU, D., ZHU, L., IGALAVITHANA, A. D., DISSANAYAKE, P. D., RINKLEBE, J. et al. (2020). Recent advances in control technologies for non-point source pollution with nitrogen and phosphorous from agricultural runoff: Current practices and future prospects. *Applied Biological Chemistry* **63**.
ZHANG, S., SHIH, Y.-C. T. and MÜLLER, P. (2007). A spatially-adjusted Bayesian additive regression tree model to merge two datasets. *Bayesian Anal.* **2** 611–633. MR2342177 https://doi.org/10.1214/07-BA224
ZIGLER, C., LIU, V., MEALLI, F. and FORASTIERE, L. (2023). Bipartite interference and air pollution transport: Estimating health effects of power plant interventions.
ZIGLER, C. M. and PAPADOGEORGOU, G. (2021). Bipartite causal inference with interference. *Statist. Sci.* **36** 109–123. MR4194206 https://doi.org/10.1214/19-STS749

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2775–2795
https://doi.org/10.1214/24-AOAS1905

A NEW MULTIPLE-MEDIATOR MODEL MAXIMALLY UNCOVERING THE MEDIATION PATHWAY: EVALUATING THE ROLE OF NEUROIMAGING MEASURES IN AGE-RELATED COGNITIVE DECLINE

BY HWIYOUNG LEE[1,a], CHIXIANG CHEN[2,c], PETER KOCHUNOV[3,d], L. ELLIOT HONG[3,e] AND SHUO CHEN[1,b]

[1]*Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore,* [a]*hwiyoung.lee@somumaryland.edu,* [b]*shuochen@som.umaryland.edu*

[2]*Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland School of Medicine,* [c]*chixiang.chen@som.umaryland.edu*

[3]*Department of Psychiatry and Behavioral Science, University of Texas Health Science Center Houston,* [d]*peter.kochunov@uth.tmc.edu,* [e]*L.Elliot.Hong@uth.tmc.edu*

Aging changes brain functions and structures in a downward trajectory and consequently leads to a decline in neurocognitive performance. Our research is motivated by understanding whether and to what extent the age-effect on cognitive decline can be explained by neuroimaging measures. We consider a new mediation model with age as an independent variable, while treating neuroimaging data and cognitive function as the multiple mediators and outcome, respectively. Given that the brain is the primary organ responsible for cognitive function, it is neurobiologically intuitive that the age-related decline in cognition is largely mediated through neuroimaging measures. Additionally, cognitive function is localized to certain regions of the brain rather than being a function of the entire brain. Taking these factors into account, we propose a novel mediation model with multiple mediators that aims to maximally uncover the mediation pathway while simultaneously identifying active neuroimaging mediators by imposing an ℓ_1 penalty and ℓ_2 constraint. We develop a computationally efficient algorithm to handle the nonconvex optimization problem of penalized mediation proportion maximization. We apply our method to a data example of 37,441 participants of UK Biobank with cortical gray-matter thickness and white-matter integrity measures and cognitive performance scores. Our results show that the mediation effect of brain-imaging variables can explain 97% of age-related cognitive decline.

REFERENCES

ANDREWS, R. M. and DIDELEZ, V. (2021). Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology* **32** 209–219.

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **51** 1173–1182. https://doi.org/10.1037//0022-3514.51.6.1173

BECK, A. (2017). *First-Order Methods in Optimization*. *MOS-Siam Series on Optimization* **25**. SIAM, Philadelphia, PA. MR3719240 https://doi.org/10.1137/1.9781611974997

BETHLEHEM, R. A. I., SEIDLITZ, J., WHITE, S. R. et al. (2022). Brain charts for the human lifespan. *Nature* **604** 525–533.

BETTIO, L. E. B., RAJENDRAN, L. and GIL-MOHAPEL, J. (2017). The effects of aging in the hippocampus and cognitive decline. *Neurosci. Biobehav. Rev.* **79** 66–86. https://doi.org/10.1016/j.neubiorev.2017.04.030

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

BUTLER, E. R., CHEN, A., RAMADAN, R., LE, T. T., RUPAREL, K., MOORE, T. M., SATTERTHWAITE, T. D., ZHANG, F., SHOU, H. et al. (2021). Pitfalls in brain age analyses. *Hum. Brain Mapp.* **42** 4092.

CHÉN, O. Y., CRAINICEANU, C., OGBURN, E. L., CAFFO, B. S., WAGER, T. D. and LINDQUIST, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19** 121–136. MR3799607 https://doi.org/10.1093/biostatistics/kxx027

Key words and phrases. Aging, maximal mediation effect, neuroimaging, optimization.

COLE, J. H. and FRANKE, K. (2017). Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci*. **40** 681–690. https://doi.org/10.1016/j.tins.2017.10.001

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

FRANKE, K. and GASER, C. (2019). Ten years of BrainAGE as a neuroimaging biomarker of brain aging: What insights have we gained?. *Front. Neurol.* **10** 789. https://doi.org/10.3389/fneur.2019.00789

GRADY, C. L. (1998). Brain imaging and age-related changes in cognition. *Exp. Gerontol.* **33** 661–673. https://doi.org/10.1016/s0531-5565(98)00022-9

GRADY, C. L. (2000). Functional brain imaging and age-related changes in cognition. *Biol. Psychol.* **54** 259–281. https://doi.org/10.1016/s0301-0511(00)00059-4

HEDDEN, T., SCHULTZ, A. P., RIECKMANN, A., MORMINO, E. C., JOHNSON, K. A., SPERLING, R. A. and BUCKNER, R. L. (2014). Multiple brain markers are linked to age-related variation in cognition. *Cereb. Cortex* **26** 1388–1400.

HOOGENDAM, Y. Y., HOFMAN, A., VAN DER GEEST, J. N., VAN DER LUGT, A. and IKRAM, M. A. (2014). Patterns of cognitive function in aging: The Rotterdam study. *Eur. J. Epidemiol.* **29** 133–140. https://doi.org/10.1007/s10654-014-9885-4

IMAI, K., KEELE, L. and TINGLEY, D. (2010). A general approach to causal mediation analysis. *Psychol. Methods* **15** 309–334. https://doi.org/10.1037/a0020761

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

KAUP, A. R., MIRZAKHANIAN, H., JESTE, D. V. and EYLER, L. T. (2011). A review of the brain structure correlates of successful cognitive aging. *J. Neuropsychiatry Clin. Neurosci.* **23** 6–15. https://doi.org/10.1176/jnp.23.1.jnp6

KOCHUNOV, P., GLAHN, D. C., LANCASTER, J., THOMPSON, P. M., KOCHUNOV, V., ROGERS, B., FOX, P., BLANGERO, J. and WILLIAMSON, D. E. (2011). Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan. *NeuroImage* **58** 41–49. https://doi.org/10.1016/j.neuroimage.2011.05.050

KOCHUNOV, P., JAHANSHAD, N., MARCUS, D., WINKLER, A., SPROOTEN, E., NICHOLS, T. E., WRIGHT, S. N., HONG, L. E., PATEL, B. et al. (2015). Heritability of fractional anisotropy in human white matter: A comparison of human connectome project and ENIGMA-DTI data. *NeuroImage* **111** 300–311.

LEE, H., CHEN, C., KOCHUNOV, P., HONG, L. E. and CHEN, S. (2024). Supplement to "A new multiple-mediator model maximally uncovering the mediation pathway: Evaluating the role of neuroimaging measures in age-related cognitive decline." https://doi.org/10.1214/24-AOAS1905SUPPA, https://doi.org/10.1214/24-AOAS1905SUPPB

LIEM, F., VAROQUAUX, G., KYNAST, J., BEYER, F., MASOULEH, S. K., HUNTENBURG, J. M., LAMPE, L., RAHIM, M., ABRAHAM, A. et al. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage* **148** 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005

LIN, Y.-C., SHIH, Y.-C., TSENG, W.-Y. I., CHU, Y.-H., WU, M.-T., CHEN, T.-F., TANG, P.-F. and CHIU, M.-J. (2014). Cingulum correlates of cognitive functions in patients with mild cognitive impairment and early Alzheimer's disease: A diffusion spectrum imaging study. *Brain Topography Volume* **27** 393–402.

LUO, Z.-Q., MA, W.-K., SO, A. M.-C., YE, Y. and ZHANG, S. (2010). Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Process. Mag.* **27** 20–34.

MACKINNON, D. P., KRULL, J. L. and LOCKWOOD, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prev. Sci.* **1** 173–181. https://doi.org/10.1023/a:1026595011371

MORRISON, J. H. and BAXTER, M. G. (2012). The ageing cortical synapse: Hallmarks and implications for cognitive decline. *Nat. Rev. Neurosci.* **13** 240–250. https://doi.org/10.1038/nrn3200

NÄSLUND, J., HAROUTUNIAN, V., MOHS, R., DAVIS, K. L., DAVIES, P., GREENGARD, P. and BUXBAUM, J. D. (2000). Correlation between elevated levels of amyloid beta-peptide in the brain and cognitive decline. *JAMA* **283** 1571–1577. https://doi.org/10.1001/jama.283.12.1571

NIU, X., ZHANG, F., KOUNIOS, J. and LIANG, H. (2020). Improved prediction of brain age using multimodal neuroimaging data. *Hum. Brain Mapp.* **41** 1626–1643.

PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

PARK, D. C. and REUTER-LORENZ, P. (2009). The adaptive brain: Aging and neurocognitive scaffolding. *Annu. Rev. Psychol.* **60** 173.

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

SALTHOUSE, T. A. (2011). Neuroanatomical substrates of age-related cognitive decline. *Psychol. Bull.* **137** 753–784. https://doi.org/10.1037/a0023262

SERANG, S., JACOBUCCI, R., BRIMHALL, K. C. and GRIMM, K. J. (2017). Exploratory mediation analysis via regularization. *Struct. Equ. Model.* **24** 733–744. MR3693630 https://doi.org/10.1080/10705511.2017.1311775

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

TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205 https://doi.org/10.1214/11-AOS878

VANDENBERGHE, L. and BOYD, S. (1996). Semidefinite programming. *SIAM Rev.* **38** 49–95. MR1379041 https://doi.org/10.1137/1038003

VANDERWEELE, T. J. and VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiol. Methods* **2** 95–115. https://doi.org/10.1515/em-2012-0010

WANG, C., HU, J., BLASER, M. J. and LI, H. (2019). Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. *Bioinformatics* **36** 347–355.

ZHANG, H., ZHENG, Y., ZHANG, Z., GAO, T., JOYCE, B., YOON, G., ZHANG, W., SCHWARTZ, J., JUST, A. et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32** 3150–3154.

ZHAO, B., LI, T., YANG, Y., WANG, X., LUO, T., SHAN, Y., ZHU, Z., XIONG, D., HAUBERG, M. E. et al. (2021). Common genetic variation influencing human white matter microstructure. *Science* **372** 1304.

ZHAO, Y., LI, L. and CAFFO, B. S. (2021). Multimodal neuroimaging data integration and pathway analysis. *Biometrics* **77** 879–889. MR4320664 https://doi.org/10.1111/biom.13351

ZHAO, Y., LI, L. and INITIATIVE, A. D. N. (2022). Multimodal data integration via mediation analysis with high-dimensional exposures and mediators. *Hum. Brain Mapp.* **43** 2519–2533.

ZHAO, Y., LINDQUIST, M. A. and CAFFO, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Comput. Statist. Data Anal.* **142** 106835. MR4001130 https://doi.org/10.1016/j.csda.2019.106835

ZHAO, Y. and LUO, X. (2022b). Pathway Lasso: Pathway estimation and selection with high-dimensional mediators. *Stat. Interface* **15** 39–50. MR4305014 https://doi.org/10.4310/21-SII673

ZIEGLER, D. A., PIGUET, O., SALAT, D. H., PRINCE, K., CONNALLY, E. and CORKIN, S. (2010). Cognition in healthy aging is related to regional white matter integrity, but not cortical thickness. *Neurobiol. Aging* **31** 1912–1926.

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

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2796–2817
https://doi.org/10.1214/24-AOAS1906

INDIVIDUAL DYNAMIC PREDICTION FOR CURE AND SURVIVAL BASED ON LONGITUDINAL BIOMARKERS

BY CAN XIE[1,a], XUELIN HUANG[1,b], RUOSHA LI[2,c], ALEXANDER TSODIKOV[3,d] AND KAPIL BHALLA[4,e]

[1]*Department of Biostatistics, University of Texas MD Anderson Cancer Center,* [a]*CXie1@mdanderson.org,* [b]*xlhuang@mdanderson.org*

[2]*Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston,* [c]*ruosha.li@uth.tmc.edu*

[3]*Department of Biostatistics, University of Michigan,* [d]*tsodikov@umich.edu*

[4]*Department of Leukemia, University of Texas MD Anderson Cancer Center,* [e]*kbhalla@mdanderson.org*

To optimize personalized treatment strategies and extend patients' survival times, it is critical to accurately predict patients' prognoses at all stages, from disease diagnosis to follow-up visits. The longitudinal biomarker measurements during visits are essential for this prediction purpose. Patients' ultimate concerns are cure and survival. However, in many situations there is no clear biomarker indicator for cure. We propose a comprehensive joint model of longitudinal and survival data and a landmark cure model, incorporating proportions of potentially cured patients. The survival distributions in the joint and landmark models are specified through flexible hazard functions with the proportional hazards as a special case, allowing other patterns such as crossing hazard and survival functions. Formulas are provided for predicting each individual's probabilities of future cure and survival at any time point based on his or her current biomarker history. Simulations show that, with these comprehensive and flexible properties, the proposed cure models outperform standard cure models in terms of predictive performance, measured by the time-dependent area under the curve of receiver operating characteristic, Brier score, and integrated Brier score. The use and advantages of the proposed models are illustrated by their application to a study of patients with chronic myeloid leukemia.

REFERENCES

AMANPOUR, F., AKBARI, S., LOOHA, M. A., ABDEHAGH, M. and POURHOSEINGHOLI, M. A. (2019). Mixture cure model for estimating short-term and long-term colorectal cancer survival. *Gastroenterology Hepatology Bed Bench* **12** S37–S43.

BARBIERI, A. and LEGRAND, C. (2020). Joint longitudinal and time-to-event cure models for the assessment of being cured. *Stat. Methods Med. Res.* **29** 1256–1270. MR4097143 https://doi.org/10.1177/0962280219853599

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc. Ser. B* **11** 15–53.

BØVELSTAD, H. M. and BORGAN, Ø. (2011). Assessment of evaluation criteria for survival prediction from genomic data. *Biom. J.* **53** 202–216. MR2897397 https://doi.org/10.1002/bimj.201000048

BROWN, E. R. and IBRAHIM, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59** 686–693. MR2004274 https://doi.org/10.1111/1541-0420.00079

GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18** 2529–2545. https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5

GRAMBSCH, P. M. and THERNEAU, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81** 515–526. MR1311094 https://doi.org/10.1093/biomet/81.3.515

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

Key words and phrases. Joint modeling, nonproportional hazard function, longitudinal biomarker, cure model, linear mixed model.

HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344.

HOLLAND, J. F. (2008). Breaking the cure barrier 25 years later. *J. Clin. Oncol.* **26** 1575. https://doi.org/10.1200/JCO.2007.13.8107

KIM, S., ZENG, D., LI, Y. and SPIEGELMAN, D. (2013). Joint modeling of longitudinal and cure-survival data. *J. Stat. Theory Pract.* **7** 324–344. MR3196603 https://doi.org/10.1080/15598608.2013.772036

LI, L. and WU, C. (2016). tdROC: Nonparametric estimation of time-dependent ROC curve from right censored survival data R package version 1.0.

LI, J., YU, T., LV, J. and LEE, M.-L. T. (2021). Semiparametric model averaging prediction for lifetime data via hazards regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 1187–1209. MR4347709 https://doi.org/10.1111/rssc.12502

LI, L., GREENE, T. and HU, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat. Methods Med. Res.* **27** 2264–2278. MR3825906 https://doi.org/10.1177/0962280216680239

MALLER, R. A. and ZHOU, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* **79** 731–739. MR1209474 https://doi.org/10.1093/biomet/79.4.731

PAN, J., BAO, Y., DAI, H. and FANG, H.-B. (2014). Joint longitudinal and survival-cure models in tumour xenograft experiments. *Stat. Med.* **33** 3229–3240. MR3260540 https://doi.org/10.1002/sim.6175

PAPAGEORGIOU, G., MAUFF, K., TOMER, A. and RIZOPOULOS, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annu. Rev. Stat. Appl.* **6** 223–240. MR3939519 https://doi.org/10.1146/annurev-statistics-030718-105048

PENG, Y. and XU, J. (2012). An extended cure model and model selection. *Lifetime Data Anal.* **18** 215–233. MR2903721 https://doi.org/10.1007/s10985-011-9213-1

SHAH, N. P., KANTARJIAN, H. M., KIM, D.-W., RÉA, D., DORLHIAC-LLACER, P. E., MILONE, J. H., VELA-OJEDA, J., SILVER, R. T., KHOURY, H. J. et al. (2008). Intermittent target inhibition with dasatinib 100 mg once daily preserves efficacy and improves tolerability in imatinib-resistant and -intolerant chronic-phase chronic myeloid leukemia. *J. Clin. Oncol.* **26** 3204–3212. https://doi.org/10.1200/JCO.2007.14.9260

SHI, B., WEI, P. and HUANG, X. (2021). Functional principal component based landmark analysis for the effects of longitudinal cholesterol profiles on the risk of coronary heart disease. *Stat. Med.* **40** 650–667. MR4198437 https://doi.org/10.1002/sim.8794

SHI, H. and YIN, G. (2017). Landmark cure rate models with time-dependent covariates. *Stat. Methods Med. Res.* **26** 2042–2054. MR3712219 https://doi.org/10.1177/0962280217708681

SONG, H., PENG, Y. and TU, D. (2012). A new approach for joint modelling of longitudinal measurements and survival times with a cure fraction. *Canad. J. Statist.* **40** 207–224. MR2927743 https://doi.org/10.1002/cjs.11127

TSODIKOV, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics* **54** 1508–1516.

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

VAN HOUWELINGEN, H. C. and PUTTER, H. (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. *Lifetime Data Anal.* **14** 447–463. MR2464769 https://doi.org/10.1007/s10985-008-9099-8

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

WANG, Y., WANG, W. and TANG, Y. (2021). A Bayesian semiparametric accelerate failure time mixture cure model. *Int. J. Biostat.* **18** 473–485.

XIE, C., HUANG, X., LI, R. and PISTERS, P. W. T. (2022). A flexible-hazards cure model with application to patients with soft tissue sarcoma. *Stat. Med.* **41** 5698–5714. MR4515037 https://doi.org/10.1002/sim.9588

XIE, C., HUANG, X., LI, R., TSODIKOV, A. and BHALLA, K. (2024). Supplement to "Individual dynamic prediction for cure and survival based on longitudinal biomarkers." https://doi.org/10.1214/24-AOAS1906SUPPA, https://doi.org/10.1214/24-AOAS1906SUPPB

YAKOVLEV, A. Y. and TSODIKOV, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. *Series in Mathematical Biology and Medicine* **1**. World Scientific, Singapore.

YANG, L., SONG, H., PENG, Y. and TU, D. (2020). Joint analysis of longitudinal measurements and survival times with a cure fraction based on partly linear mixed and semiparametric cure models. *Pharm. Stat.* **20** 362–374.

YILMAZ, Y. E., LAWLESS, J. F., ANDRULIS, I. L. and BULL, S. B. (2013). Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *J. Clin. Oncol.* **31** 2047–2054. https://doi.org/10.1200/JCO.2012.46.6615

YU, M., TAYLOR, J. M. G. and SANDLER, H. M. (2008). Individual prediction in prostate cancer studies using a joint longitudinal survival-cure model. *J. Amer. Statist. Assoc.* **103** 178–187. MR2420225 https://doi.org/10.1198/016214507000000400

ZHANG, Y., HAN, X. and SHAO, Y. (2021). The ROC of Cox proportional hazards cure models with application in cancer studies. *Lifetime Data Anal.* **27** 195–215. MR4227847 https://doi.org/10.1007/s10985-021-09516-6

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

ZHU, Y., HUANG, X. and LI, L. (2020). Dynamic prediction of time to a clinical event with sparse and irregularly measured longitudinal biomarkers. *Biom. J.* **62** 1371–1393. MR4164143 https://doi.org/10.1002/bimj.201900112

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2818–2839
https://doi.org/10.1214/24-AOAS1907

NEURAL NETWORKS FOR EXTREME QUANTILE REGRESSION WITH AN APPLICATION TO FORECASTING OF FLOOD RISK

BY OLIVIER C. PASCHE[a] AND SEBASTIAN ENGELKE[b]

Research Center for Statistics, University of Geneva, [a]*olivier.pasche@unige.ch,* [b]*sebastian.engelke@unige.ch*

Risk assessment for extreme events requires accurate estimation of high quantiles that go beyond the range of historical observations. When the risk depends on the values of observed predictors, regression techniques are used to interpolate in the predictor space. We propose the EQRN model that combines tools from neural networks and extreme value theory into a method capable of extrapolation in the presence of complex predictor dependence. Neural networks can naturally incorporate additional structure in the data. We develop a recurrent version of EQRN that is able to capture complex sequential dependence in time series. We apply this method to forecast flood risk in the Swiss Aare catchment. It exploits information from multiple covariates in space and time to provide one-day-ahead predictions of return levels and exceedance probabilities. This output complements the static return level from a traditional extreme value analysis, and the predictions are able to adapt to distributional shifts as experienced in a changing climate. Our model can help authorities to manage flooding more effectively and to minimize their disastrous impacts through early warning systems.

REFERENCES

ANDRES, N., STEEB, N., BADOUX, A. and HEGG, C., eds. (2021). Extremhochwasser an der Aare. Hauptbericht Projekt EXAR. Methodik und Resultate. [Extreme flooding of the Aare. Main Report on the EXAR Project. Methodology and results.]. WSL Berichte 104.

ASADI, P., DAVISON, A. C. and ENGELKE, S. (2015). Extremes on river networks. *Ann. Appl. Stat.* **9** 2023–2050. MR3456363 https://doi.org/10.1214/15-AOAS863

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

BALKEMA, A. A. and DE HAAN, L. (1974). Residual life time at great age. *Ann. Probab.* **2** 792–804. MR0359049 https://doi.org/10.1214/aop/1176996548

BEZZOLA, G. R. and HEGG, C. (2007). Ereignisanalyse Hochwasser 2005, Teil 1—Prozesse, Schäden und erste Einordnung [Event analysis of the 2005 flood, Part 1—Processes, damage and initial classification]. Technical report, Federal Office for the Environment FOEN, Swiss Federal Institute for Forest, Snow and Landscape Research WSL. Umwelt-Wissen Nr. 0707. 215 S.

BOULAGUIEM, Y., ZSCHEISCHLER, J., VIGNOTTO, E., VAN DER WIEL, K. and ENGELKE, S. (2022). Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environ. Data Sci.* **1** e5.

BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64.

BÜCHER, A. and ZHOU, C. (2021). A horse race between the block maxima method and the peak-over-threshold approach. *Statist. Sci.* **36** 360–378. MR4293095 https://doi.org/10.1214/20-STS795

CANNON, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.* **37** 1277–1284.

CANNON, A. J. (2012). Neural networks for probabilistic environmental prediction: Conditional Density Estimation Network Creation and Evaluation (CaDENCE) in R. *Comput. Geosci.* **41** 126–135.

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

CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160 https://doi.org/10.1214/009053604000001165

Key words and phrases. Extreme value theory, generalized Pareto distribution, machine learning, prediction, recurrent neural network.

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. and BENGIO, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. Conf. Empir. Methods Nat. Lang. Process* 1724–1734.

CLEVELAND, R. B., CLEVELAND, W. S., MCRAE, J. E. and TERPENNING, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *J. Off. Stat.* **6** 3–73.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. With a discussion. MR0893334

DAOUIA, A., GARDES, L., GIRARD, S. and LEKINA, A. (2011). Kernel estimators of extreme level curves. *TEST* **20** 311–333. MR2834049 https://doi.org/10.1007/s11749-010-0196-0

DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. MR2825422

ELMAN, J. L. (1990). Finding structure in time. *Cogn. Sci.* **14** 179–211.

ENGELKE, S. and HITZ, A. S. (2020). Graphical models for extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 871–932. With discussions. MR4136498

FISCHER, E. M., SIPPEL, S. and KNUTTI, R. (2021). Increasing probability of record-shattering climate extremes. *Nat. Clim. Change* **11** 689–695.

FISHER, R. A. and TIPPETT, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Math. Proc. Camb. Philos. Soc.* **24** 180–190. Cambridge Univ. Press, Cambridge.

GARDES, L. and STUPFLER, G. (2019). An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT* **17** 109–144. MR3916311 https://doi.org/10.1007/s10687-013-0174-5

GERS, F. A., SCHMIDHUBER, J. and CUMMINS, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Comput.* **12** 2451–2471. https://doi.org/10.1162/089976600300015015

GERS, F. A., SCHRAUDOLPH, N. N. and SCHMIDHUBER, J. (2003). Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **3** 115–143. MR1966056 https://doi.org/10.1162/153244303768966139

GNECCO, N., TEREFE, E. M. and ENGELKE, S. (2022). Extremal random forests. Available at arXiv:2201.12865.

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

HOCHREITER, S. and SCHMIDHUBER, J. (1997). Long short-term memory. *Neural Comput.* **9** 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4** 251–257.

JOZEFOWICZ, R., ZAREMBA, W. and SUTSKEVER, I. (2015). An empirical exploration of recurrent network architectures. In *Proc. 32nd Int. Conf. Mach. Learn.* **37** 2342–2350.

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour.* **25** 1287–1304.

KEEF, C., TAWN, J. and SVENSSON, C. (2009). Spatial risk assessment for extreme river flows. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 601–618. MR2750258 https://doi.org/10.1111/j.1467-9876.2009.00672.x

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. In *3rd Int. Conf. Learn. Repres.*

KINSVATER, P., FRIED, R. and LILIENTHAL, J. (2016). Regional extreme value index estimation and a test of tail homogeneity. *Environmetrics* **27** 103–115. MR3481322 https://doi.org/10.1002/env.2376

KLAMBAUER, G., UNTERTHINER, T., MAYR, A. and HOCHREITER, S. (2017). Self-normalizing neural networks. In *Proc. 31st Int. Conf. Neural Inf. Process. Syst* 972–981.

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

KOH, J. (2023). Gradient boosting with extreme-value theory for wildfire prediction. *Extremes* **26** 273–299. MR4577408 https://doi.org/10.1007/s10687-022-00454-6

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444. https://doi.org/10.1038/nature14539

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

LI, D. and WANG, H. J. (2019). Extreme quantile estimation for autoregressive models. *J. Bus. Econom. Statist.* **37** 661–670. MR4016161 https://doi.org/10.1080/07350015.2017.1408469

LINDSTRÖM, G., JOHANSSON, B., PERSSON, M., GARDELIN, M. and BERGSTRÖM, S. (1997). Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **201** 272–288.

LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Adv. Neural Inf. Process. Syst* **30**.

MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394

PICKANDS, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131. MR0423667
ROMANO, Y., PATTERSON, E. and CANDES, E. (2019). Conformalized quantile regression. In *Adv. Neural Inf. Process. Syst* **32**.
SCARSELLI, F., GORI, M., TSOI, A. C., HAGENBUCHNER, M. and MONFARDINI, G. (2009). The graph neural network model. *IEEE Trans. Neural Netw.* **20** 61–80. https://doi.org/10.1109/TNN.2008.2005605
SCHMOCKER-FACKEL, P. and NAEF, F. (2010). Changes in flood frequencies in Switzerland since 1500. *Hydrol. Earth Syst. Sci.* **14** 1581–1594.
SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958. MR3231592
TIELEMAN, T. and HINTON, G. (2012). Lecture 6.5—RMSProp Technical report. COURSERA: Neural networks for machine learning.
VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U. and POLOSUKHIN, I. (2017). Attention is all you need. In *Adv. Neural Inf. Process. Syst* **30**.
VELTHOEN, J., CAI, J.-J., JONGBLOED, G. and SCHMEITS, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes* **22** 599–622. MR4031851 https://doi.org/10.1007/s10687-019-00355-1
VELTHOEN, J., DOMBRY, C., CAI, J.-J. and ENGELKE, S. (2023). Gradient boosting for extreme quantile regression. *Extremes* **26** 639–667. MR4669002 https://doi.org/10.1007/s10687-023-00473-x
WANG, H. J., LI, D. and HE, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *J. Amer. Statist. Assoc.* **107** 1453–1464. MR3036407 https://doi.org/10.1080/01621459.2012.716382
WERBOS, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw.* **1** 339–356.
WOLPERT, D. H. (1992). Stacked generalization. *Neural Netw.* **5** 241–259.
WU, Z., PAN, S., CHEN, F., LONG, G., ZHANG, C. and YU, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32** 4–24. MR4205495 https://doi.org/10.1109/tnnls.2020.2978386
YOUNGMAN, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *J. Amer. Statist. Assoc.* **114** 1865–1879. MR4047306 https://doi.org/10.1080/01621459.2018.1529596
PASCHE, O. C. and ENGELKE, S. (2024). Supplement to "Neural networks for extreme quantile regression with an application to forecasting of flood risk." https://doi.org/10.1214/24-AOAS1907SUPPA, https://doi.org/10.1214/24-AOAS1907SUPPB
ZHANG, W., QUAN, H. and SRINIVASAN, D. (2019). An improved quantile regression neural network for probabilistic load forecasting. *IEEE Trans. Smart Grid* **10** 4425–4434.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2840–2862
https://doi.org/10.1214/24-AOAS1908

IMPLICIT GENERATIVE PRIOR FOR BAYESIAN NEURAL NETWORKS

BY YIJIA LIU[a] AND XIAO WANG[b]

Department of Statistics, Purdue University, [a]*liu2300@purdue.edu,* [b]*wangxiao@purdue.edu*

Predictive uncertainty quantification is crucial for reliable decision-making in various applied domains. Bayesian neural networks offer a powerful framework for this task. However, defining meaningful priors and ensuring computational efficiency remain significant challenges, especially for complex real-world applications. This paper addresses these challenges by proposing a novel neural adaptive empirical Bayes (NA-EB) framework. NA-EB leverages a class of implicit generative priors derived from low-dimensional distributions. This allows for efficient handling of complex data structures and effective capture of underlying relationships in real-world datasets. The proposed NA-EB framework combines variational inference with a gradient ascent algorithm. This enables simultaneous hyperparameter selection and approximation of the posterior distribution, leading to improved computational efficiency. We establish the theoretical foundation of the framework through posterior and classification consistency. We demonstrate the practical applications of our framework through extensive evaluations on a variety of tasks, including the two-spiral problem, regression, 10 UCI datasets, and image classification tasks on both MNIST and CIFAR-10 datasets. The results of our experiments highlight the superiority of our proposed framework over existing methods, such as sparse variational Bayesian and generative models, in terms of prediction accuracy and uncertainty quantification.

REFERENCES

ATANOV, A., ASHUKHA, A., STRUMINSKY, K., VETROV, D. and WELLING, M. (2018). The deep weight prior. arXiv preprint. Available at arXiv:1810.06943.

ATCHADÉ, Y. F. (2011). A computational framework for empirical Bayes inference. *Stat. Comput.* **21** 463–473. MR2826685 https://doi.org/10.1007/s11222-010-9182-3

BAI, J., SONG, Q. and CHENG, G. (2020). Efficient variational inference for sparse deep learning with theoretical guarantee. *Adv. Neural Inf. Process. Syst.* **33** 466–476.

BASU, S., KARKI, M., GANGULY, S., DIBIANO, R., MUKHOPADHYAY, S., GAYAKA, S., KANNAN, R. and NEMANI, R. (2017). Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Process. Lett.* **45** 855–867.

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

BHATTACHARYA, S., LIU, Z. and MAITI, T. (2020). Variational bayes neural network: Posterior consistency, classification accuracy and computational challenges. arXiv preprint. Available at arXiv:2011.09592.

BHATTACHARYA, S. and MAITI, T. (2021). Statistical foundation of variational Bayes neural networks. *Neural Netw.* **137** 151–173. https://doi.org/10.1016/j.neunet.2021.01.027

BISHOP, C. M. (1997). Bayesian neural networks. *J. Braz. Comput. Soc.* **4** 61–68.

BLEI, D. M. and LAFFERTY, J. D. (2007). A correlated topic model of science. *Ann. Appl. Stat.* **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

BLUNDELL, C., CORNEBISE, J., KAVUKCUOGLU, K. and WIERSTRA, D. (2015). Weight uncertainty in neural network. In *International Conference on Machine Learning*. 1613–1622. PMLR.

CARLIN, B. P. and LOUIS, T. A. (2009). *Bayesian Methods for Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2442364

Key words and phrases. Deep neural networks, empirical Bayes, latent variable model, stochastic gradient method, variational inference.

CHEN, Y., GAO, Q. and WANG, X. (2022). Inferential Wasserstein generative adversarial networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 83–113. MR4400391 https://doi.org/10.1111/rssb.12476

CHING, T., HIMMELSTEIN, D. S., BEAULIEU-JONES, B. K., KALININ, A. A., DO, B. T., WAY, G. P., FERRERO, E., AGAPOW, P.-M., ZIETZ, M. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15** 20170387.

DUA, D. and GRAFF, C. (2017). UCI machine learning repository.

DUSENBERRY, M. W., JERFEL, G., WEN, Y., MA, Y.-A., SNOEK, J., HELLER, K., LAKSHMINARAYANAN, B. and TRAN, D. (2020). Efficient and scalable Bayesian neural nets with rank-1 factors. In *Proceedings of the* 37*th International Conference on Machine Learning*. *ICML*'20. JMLR.org.

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

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

GAL, Y. and GHAHRAMANI, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*. 1050–1059. PMLR.

GHOSH, S., YAO, J. and DOSHI-VELEZ, F. (2019). Model selection in Bayesian neural networks via horseshoe priors. *J. Mach. Learn. Res.* **20** Paper No. 182, 46. MR4048993

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**.

GRAVES, A. (2011). Practical variational inference for neural networks. *Adv. Neural Inf. Process. Syst.* **24**.

HAN, X., ZHENG, H. and ZHOU, M. (2022). CARD: Classification and regression diffusion models. arXiv preprint. Available at arXiv:2206.07275.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. 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

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HERNÁNDEZ-LOBATO, J. M. and ADAMS, R. P. (2015a). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning* **37**. *ICML*'15. 1861–1869. JMLR.org.

HERNÁNDEZ-LOBATO, J. M. and ADAMS, R. P. (2015a). Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*. 1861–1869. PMLR.

HINTON, G. E. and VAN CAMP, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory* 5–13.

HOFFMAN, J., ROBERTS, D. A. and YAIDA, S. (2019). Robust learning with jacobian regularization. arXiv preprint. Available at arXiv:1908.02729.

HUBIN, A., STORVIK, G. and FROMMLET, F. (2018). Deep Bayesian regression models. arXiv preprint. Available at arXiv:1806.02160.

IMMER, A., BAUER, M., FORTUIN, V., RÄTSCH, G. and EMTIYAZ, K. M. (2021). Scalable marginal likelihood estimation for model selection in deep learning. In *International Conference on Machine Learning* 4563–4573. PMLR.

IZMAILOV, P., VIKRAM, S., HOFFMAN, M. D. and WILSON, A. G. G. (2021). What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning* 4629–4640. PMLR.

JAVID, K., HANDLEY, W., HOBSON, M. and LASENBY, A. (2020). Compromise-free Bayesian neural networks. arXiv preprint. Available at arXiv:2004.12211.

KAMNITSAS, K., LEDIG, C., NEWCOMBE, V. F. J., SIMPSON, J. P., KANE, A. D., MENON, D. K., RUECKERT, D. and GLOCKER, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36** 61–78. https://doi.org/10.1016/j.media.2016.10.004

KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational bayes. arXiv preprint. Available at arXiv:1312.6114.

LAKSHMINARAYANAN, B., PRITZEL, A. and BLUNDELL, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **30**.

LAMPINEN, J. and VEHTARI, A. (2001). Bayesian approach for neural networks–review and case studies. *Neural Netw.* **14** 257–274. https://doi.org/10.1016/s0893-6080(00)00098-8

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.

LEIBIG, C., ALLKEN, V., AYHAN, M. S., BERENS, P. and WAHL, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Sci. Rep.* **7** 17816. https://doi.org/10.1038/s41598-017-17876-z

LIU, Y. and WANG, X. (2024). Supplement to "Implicit generative prior for Bayesian neural networks." https://doi.org/10.1214/24-AOAS1908SUPPA, https://doi.org/10.1214/24-AOAS1908SUPPB

LOUIZOS, C., ULLRICH, K. and WELLING, M. (2017). Bayesian compression for deep learning. In *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. *NIPS'*17. 3290–3300. Curran Associates, Red Hook, NY, USA.

MACKAY, D. J. (1995). Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. *Netw. Comput. Neural Syst.* **6** 469.

MOLCHANOV, D., ASHUKHA, A. and VETROV, D. (2017). Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*. 2498–2507. PMLR.

MULLACHERY, V., KHERA, A. and HUSAIN, A. (2018). Bayesian neural networks. arXiv preprint. Available at arXiv:1801.07710.

NEAL, R. M. (1992). Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical report, Citeseer.

NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York.

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**.

QUINONERO-CANDELA, J., RASMUSSEN, C. E., SINZ, F., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Evaluating predictive uncertainty challenge. In *Machine Learning Challenges Workshop* 1–27. Springer, Berlin.

RANGANATH, R., GERRISH, S. and BLEI, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics* 814–822. PMLR.

ROBBINS, H. E. (1992). An empirical Bayes approach to statistics. In *Breakthroughs in Statistics* 388–394. Springer, Berlin.

RUMELHART, D. E., HINTON, G. E. and WILLIAMS, R. J. (1986). Learning representations by back-propagating errors. *Nature* **323** 533–536.

SIMARD, P. Y., STEINKRAUS, D., PLATT, J. C. et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar* 3, Edinburgh.

SPRINGENBERG, J. T., KLEIN, A., FALKNER, A. and HUTTER, F. (2016). Bayesian optimization with robust Bayesian neural networks. In *NeurIPS*.

SUN, S., CHEN, C. and CARIN, L. (2017). Learning structured weight uncertainty in Bayesian neural networks. In *Artificial Intelligence and Statistics* 1283–1292. PMLR.

TOMCZAK, M., SWAROOP, S., FOONG, A. and TURNER, R. (2021). Collapsed variational bounds for Bayesian neural networks. *Adv. Neural Inf. Process. Syst.* **34** 25412–25426.

WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*.

WENZEL, F., ROTH, K., VEELING, B. S., SWIATKOWSKI, J., TRAN, L., MANDT, S., SNOEK, J., SALIMANS, T., JENATTON, R. et al. (2020). How good is the bayes posterior in deep neural networks really? arXiv preprint. Available at arXiv:2002.02405.

WILSON, A. G. and IZMAILOV, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *Adv. Neural Inf. Process. Syst.* **33** 4697–4708.

WORRALL, D. E., WILSON, C. M. and BROSTOW, G. J. (2016). Automated retinopathy of prematurity case detection with convolutional neural networks. In *International Workshop on Deep Learning in Medical Image Analysis* 68–76. Springer, Berlin.

ZHANG, G., SUN, S., DUVENAUD, D. and GROSSE, R. (2018). Noisy natural gradient as variational inference. In *International Conference on Machine Learning* 5852–5861. PMLR, Stockholm, Sweden.

ZHOU, X., JIAO, Y., LIU, J. and HUANG, J. (2023). A deep generative approach to conditional sampling. *J. Amer. Statist. Assoc.* **118** 1837–1848. MR4646610 https://doi.org/10.1080/01621459.2021.2016424

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2863–2886
https://doi.org/10.1214/24-AOAS1909

INCORPORATING AUXILIARY INFORMATION FOR IMPROVED STATISTICAL INFERENCE AND ITS EXTENSIONS TO DISTRIBUTED ALGORITHMS WITH AN APPLICATION TO PERSONAL CREDIT

BY MIAOMIAO YU[1,a], ZHONGFENG JIANG[2,c], JIAXUAN LI[3,d] AND YONG ZHOU[1,b]

[1]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science (MOE), School of Statistics and Academy of Statistics and Interdisciplinary Sciences, East China Normal University,* [a]*mmyu@fem.ecnu.edu.cn,* [b]*yzhou@fem.ecnu.edu.cn*

[2]*Academy of Mathematics and System Sciences, Chinese Academy of Science,* [c]*jiangzhongfeng17@163.com*

[3]*Chengdu No. 7 High School,* [d]*jiaxuanlee0719@163.com*

Personal credits have always been a hot topic in the society. Among all of them, the evaluation of default risk is particularly concerned since robust estimation, based on personal information, can both help needy individuals to get loans and financial institutions to avoid losses. So far, there have been no good solutions due to limited data, especially default information. With the advent of the era of big data, it is possible to improve the effectiveness of estimates by using auxiliary information from external studies or public domains. However, the individual-level data can not be gained directly because of the emphasis on data privacy; that is, only some summarized statistics with auxiliary information are allowed to be shared. To effectively utilize external integrated auxiliary information to improve the accuracy of default risk estimation, this paper introduces a unified auxiliary information framework, which is referred as enhanced GEE method, to effectively incorporate various external summary results by employing the generalized estimating equations (GEE) approach and augmenting a weighted logarithm of confidence density on GEE function. We establish asymptotic properties for the new method and prove that it can achieve the gain of statistical efficiency compared to the study-specific estimator without any auxiliary information. Besides, a low-cost Map-Reduce procedure for the distributed statistical inference of enhanced GEE method in big data is developed that can achieve the same efficiency as the oracle enhanced GEE approach under mild condition. This method is demonstrated by an application to predict the loan default risk of bank customers in Shanghai and shown to be more effective and reliable compared with the method based on the own data only. Furthermore, the superiorities of our approach, especially the construction of the tighter confidence intervals, are also illustrated with extensive simulation studies and a real personal default risk case.

REFERENCES

ANDREWS, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica* **55** 1465–1471. MR0923471 https://doi.org/10.2307/1913568

BI, Q., WU, Y., MEI, S., YE, C., ZOU, X., ZHANG, Z., LIU, X., WEI, L., TRUELOVE, S. A. et al. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **20** 911–919.

CHATTERJEE, N., CHEN, Y.-H., MAAS, P. and CARROLL, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.* **111** 107–117. MR3494641 https://doi.org/10.1080/01621459.2015.1123157

CLAGGETT, B., XIE, M. and TIAN, L. (2014). Meta-analysis with fixed, unknown, study-specific parameters. *J. Amer. Statist. Assoc.* **109** 1660–1671. MR3293618 https://doi.org/10.1080/01621459.2014.957288

Key words and phrases. External auxiliary information, individual-level data, generalized estimating equations, confidence density, distributed statistical inference.

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
FAN, J., GUO, Y. and WANG, K. (2023). Communication-efficient accurate statistical estimation. *J. Amer. Statist. Assoc.* **118** 1000–1010. MR4595472 https://doi.org/10.1080/01621459.2021.1969238
FAN, J., WANG, D., WANG, K. and ZHU, Z. (2019). Distributed estimation of principal eigenspaces. *Ann. Statist.* **47** 3009–3031. MR4025733 https://doi.org/10.1214/18-AOS1713
FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. *Texts in Statistical Science Series*. CRC Press, London. MR1699953 https://doi.org/10.1007/978-1-4899-4549-5
FOSDICK, B. K., DEYOREO, M. and REITER, J. P. (2016). Categorical data fusion using auxiliary information. *Ann. Appl. Stat.* **10** 1907–1929. MR3592042 https://doi.org/10.1214/16-AOAS925
GOGNA, A. and MAJUMDAR, A. (2015). Matrix completion incorporating auxiliary information for recommender system design. *Expert Syst. Appl.* **42** 5789–5799.
HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123 https://doi.org/10.2307/1912775
HUANG, C.-Y. and QIN, J. (2020). A unified approach for synthesizing population-level covariate effect information in semiparametric estimation with survival data. *Stat. Med.* **39** 1573–1590. MR4098508 https://doi.org/10.1002/sim.8499
HUANG, C.-Y., QIN, J. and TSAI, H.-T. (2016). Efficient estimation of the Cox model with auxiliary subgroup survival information. *J. Amer. Statist. Assoc.* **111** 787–799. MR3538705 https://doi.org/10.1080/01621459.2015.1044090
JIANG, Z., YANG, B., QIN, J. and ZHOU, Y. (2021). Enhanced empirical likelihood estimation of incubation period of COVID-19 by integrating published information. *Stat. Med.* **40** 4252–4268. MR4300084 https://doi.org/10.1002/sim.9026
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
KUNDU, P., TANG, R. and CHATTERJEE, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106** 567–585. MR3992390 https://doi.org/10.1093/biomet/asz030
LIANG, K. (2019). Empirical Bayes analysis of RNA sequencing experiments with auxiliary information. *Ann. Appl. Stat.* **13** 2452–2482. MR4037437 https://doi.org/10.1214/19-aoas1270
LIN, D. Y. and ZENG, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* **97** 321–332. MR2650741 https://doi.org/10.1093/biomet/asq006
LIN, N. and XI, R. (2011). Aggregated estimating equation estimation. *Stat. Interface* **4** 73–83. MR2775250 https://doi.org/10.4310/SII.2011.v4.n1.a8
LIU, D., LIU, R. Y. and XIE, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *J. Amer. Statist. Assoc.* **110** 326–340. MR3338506 https://doi.org/10.1080/01621459.2014.899235
NEWEY, W. K. and MCFADDEN, D. (1994). Chapter 36 large sample estimation and hypothesis testing. *Handb. Econom.* **4** 2111–2245.
QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085 https://doi.org/10.1214/aos/1176325370
QIN, J., ZHANG, H., LI, P., ALBANES, D. and YU, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102** 169–180. MR3335103 https://doi.org/10.1093/biomet/asu048
SHEN, J., LIU, R. Y. and XIE, M. (2020). iFusion: Individualized fusion learning. *J. Amer. Statist. Assoc.* **115** 1251–1267. MR4143463 https://doi.org/10.1080/01621459.2019.1672557
SINGH, K., XIE, M. and STRAWDERMAN, W. E. (2005). Combining information from independent sources through confidence distributions. *Ann. Statist.* **33** 159–183. MR2157800 https://doi.org/10.1214/009053604000001084
SUTTON, A. J. and HIGGINS, J. P. T. (2008). Recent developments in meta-analysis. *Stat. Med.* **27** 625–650. MR2418504 https://doi.org/10.1002/sim.2934
VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2
WANG, Z., WU, C., YU, M. and TSUNG, F. (2022). Self-starting process monitoring based on transfer learning. *J. Qual. Technol.* **54** 589–604.
WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526
XIE, M., SINGH, K. and STRAWDERMAN, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106** 320–333. MR2816724 https://doi.org/10.1198/jasa.2011.tm09803

YU, M., JIANG, Z., LI, J. and ZHOU, Y. (2024). Supplement to "Incorporating auxiliary information for improved statistical inference and its extensions to distributed algorithms with an application to personal credit." https://doi.org/10.1214/24-AOAS1909SUPPA, https://doi.org/10.1214/24-AOAS1909SUPPB

ZHAN, X. and GHOSH, D. (2015). Incorporating auxiliary information for improved prediction using combination of kernel machines. *Stat. Methodol.* **22** 47–57. MR3261596 https://doi.org/10.1016/j.stamet.2014.08.001

ZHANG, H., DENG, L., SCHIFFMAN, M., QIN, J. and YU, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107** 689–703. MR4138984 https://doi.org/10.1093/biomet/asaa014

ZHANG, Y., DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* **14** 3321–3363. MR3144464

ZHOU, Y., WAN, A. T. K. and YUAN, Y. (2011). Combining least-squares and quantile regressions. *J. Statist. Plann. Inference* **141** 3814–3828. MR2823652 https://doi.org/10.1016/j.jspi.2011.06.018

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2887–2908
https://doi.org/10.1214/24-AOAS1910

EARLY EFFECTS OF 2014 U.S. MEDICAID EXPANSIONS ON MORTALITY: DESIGN-BASED INFERENCE FOR IMPACTS ON SMALL SUBGROUPS DESPITE SMALL-CELL SUPPRESSION

BY CHARLOTTE Z. MANN[1,a], BEN B. HANSEN[1,b] AND LAUREN GAYDOSH[2,c]

[1]*Department of Statistics, University of Michigan,* [a]*manncz@umich.edu,* [b]*bbh@umich.edu*
[2]*Department of Sociology, University of Texas at Austin,* [c]*lauren.gaydosh@austin.utexas.edu*

Since 2014, states in the U.S. can choose whether to adopt Medicaid expansion as part of the Affordable Care Act (ACA), relaxing eligibility requirements. This heterogeneity in policy adoption between states raises the question—would there be a difference in health outcomes for states that have not expanded insurance access if they did expand Medicaid eligibility? In this study we estimate the effect of ACA Medicaid expansion on county-level all-cause mortality in the U.S. in 2014 overall and for subgroups relevant to the racial politics surrounding the ACA. We bring a causal approach to this challenge which emphasizes observational study design, including prespecifying all analyses, matching counties on pretreatment covariates, and employing design-based inference.

A challenge facing analyses like this one is gaining access to mortality outcomes, as statistical agencies in the U.S. and elsewhere suppress cell counts of 10 or fewer in public use data. We develop a rank-sum test statistic accommodating outcomes that are coarsened in this way and that lends itself to design-based inference with county-aggregated data. As applied to impact analysis of the ACA's Medicaid expansion, the proposed method's inferences from coarsened, publicly available data are substantively the same as those that would be drawn from the complete, restricted-access data.

REFERENCES

ALTHOFF, K. N., LEIFHEIT, K. M., PARK, J. N., CHANDRAN, A. and SHERMAN, S. G. (2020). Opioid-related overdose mortality in the era of fentanyl: Monitoring a shifting epidemic by person, place, and time. *Drug Alcohol Depend*. **216** 108321.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*: *Survey and Advances*. Princeton Univ. Press, Princeton, NJ. MR0331595

ARAUJO, M. C., MARTINEZ, M. A., MARTÍNEZ, S., PÉREZ, M. and SÁNCHEZ, M. (2017). Study proposal: Do larger school grants improve educational attainment? Evidence from urban Mexico. Technical Report, Inter-American Development Bank.

ARAUJO, M. C., MARTINEZ, M. A., MARTINEZ, S., PEREZ, M. and SANCHEZ, M. (2021). Do larger school grants improve educational attainment? Evidence from urban Mexico. *Journal of Development Effectiveness* **13** 405–423.

AUSTIN, A. E., NAUMANN, R. B. and SHORT, N. A. (2021). Association between medicaid expansion and suicide mortality among nonelderly US adults. *Amer. J. Epidemiol*. **190** 1760–1769. https://doi.org/10.1093/aje/kwab130

AUSTIN, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat*. **10** 150–161. https://doi.org/10.1002/pst.433

BARNES, J. M., JOHNSON, K. J., ADJEI BOAKYE, E., SCHAPIRA, L., AKINYEMIJU, T., PARK, E. M., GRABOYES, E. M. and OSAZUWA-PETERS, N. (2021). Early medicaid expansion and cancer mortality. *J. Natl. Cancer Inst*. **113** 1714–1722.

BENITEZ, J. A. and SEIBER, E. E. (2018). US health care reform and rural America: Results from the ACA's medicaid expansions. *J. Rural Health* **34** 213–222. https://doi.org/10.1111/jrh.12284

Key words and phrases. Causal inference, observational studies, propensity score matching, rank test, partial ordering, Affordable Care Act, CDC WONDER.

BLACK, B., HOLLINGSWORTH, A., NUNES, L. and SIMON, K. (2019). The effect of health insurance on mortality: Power analysis and what we can learn from the affordable care act coverage expansions. Working Paper No. 25568, National Bureau of Economic Research.
BOMMERSBACH, T. J., ROSENHECK, R. A. and EVERETT, A. S. (2022). Suicide hot spots: Leveraging county-level data and local agencies to target prevention in high-risk areas. *Public Health Rep*. **137** 408–413. https://doi.org/10.1177/00333549211016606
BORGSCHULTE, M. and VOGLER, J. (2020). Did the ACA medicaid expansion save lives? *J. Health Econ*. **72** 102333. https://doi.org/10.1016/j.jhealeco.2020.102333
CDC (2020). Restricted-use vital statistics data. Available at https://www.cdc.gov/nchs/nvss/nvss-restricted-data.htm.
CDC (2022). Underlying cause of death 1999-2020. Available at https://wonder.cdc.gov/wonder/help/ucd.html. Date Accessed: 3/28/2022.
CDC (2023). CDC WONDER. Available at https://wonder.cdc.gov/. Date Accessed: 3/20/2023.
COCHRAN, W. G. (1972). Observational studies. In *Statistical Papers in Honor of George W. Snedecor* 77–90. Iowa State Univ. Press, Ames, IA. MR0451569
COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā Ser. A* **35** 417–446.
DALEN, J. E., WATERBROOK, K. and ALPERT, J. S. (2015). Why do so many Americans oppose the affordable care act? *Am. J. Med*. **128** 807–810. https://doi.org/10.1016/j.amjmed.2015.01.032
ELIASON, E. L. (2020). Adoption of medicaid expansion is associated with lower maternal mortality. *Womens Health Issues* **30** 147–152. https://doi.org/10.1016/j.whi.2020.01.005
FISHER, R. A. (1935). *Design of Experiments*. Oliver and Boyd, Edinburgh.
FRANZ, B., MILNER, A. N. and BROWN, R. K. (2021). Opposition to the affordable care act has little to do with health care. *Race Soc. Probl*. **13** 161–169.
FREDRICKSON, M. M., ERRICKSON, J. and HANSEN, B. B. (2020). Comment: Matching methods for observational studies derived from large administrative databases [MR4148206]. *Statist. Sci*. **35** 361–366. MR4148208 https://doi.org/10.1214/19-STS740
GIBBS, C. P., ELMORE, R. and FOSDICK, B. K. (2022). The causal effect of a timeout at stopping an opposing run in the NBA. *Ann. Appl. Stat*. **16** 1359–1379. MR4455884 https://doi.org/10.1214/21-aoas1545
GKIOULEKA, A., HUIJTS, T., BECKFIELD, J. and BAMBRA, C. (2018). Understanding the micro and macro politics of health: Inequalities, intersectionality & institutions—a research agenda. *Soc. Sci. Med*. **200** 92–98.
GOLDIN, J., LURIE, I. Z. and MCCUBBIN, J. (2020). Health insurance and mortality: Experimental evidence from taxpayer outreach. *Q. J. Econ*. **136** 1–49.
GUTH, M., GARFIELD, R. and RUDOWITZ, R. (2020). The effects of medicaid expansion under the ACA: Updated findings from a literature review.
HANSEN, B. B. (2011). Propensity score matching to extract latent experiments from nonexperimental data: A case study. In *Looking Back*: *Proceedings of a Conference in Honor of Paul W. Holland* (N. Dorans and S. Sinharay, eds.) 149–181. Springer, Berlin.
HANSEN, B. B. (2023). Matching calipers and the precision of index estimation. Preprint. Available at arXiv:2301.04109.
HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci*. **23** 219–236. MR2516821 https://doi.org/10.1214/08-STS254
HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist*. **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047
HANSEN, B. B., ROSENBAUM, P. R. and SMALL, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *J. Amer. Statist. Assoc*. **109** 133–144. MR3180552 https://doi.org/10.1080/01621459.2013.863157
HANSEN, B. B. and SALES, A. C. (2015). Comments on 'observational studies,' by William G. Cochran. *Obs. Stud*. **1** 184–193.
HEALTH RESOURCES & SERVICES ADMINISTRATION (2019). Area health resources files. Available at https://data.hrsa.gov/data/download. Date Accessed: 7/7/2020.
HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat*. **34** 598–611. MR0152070 https://doi.org/10.1214/aoms/1177704172
HOTHORN, T., BRETZ, F. and WESTFALL, P. (2008). Simultaneous inference in general parametric models. *Biom. J*. **50** 346–363. MR2521547 https://doi.org/10.1002/bimj.200810425
IHME (2013a). United States Physical Activity and Obesity Prevalence by County 2001–2011. Available at http://ghdx.healthdata.org/record/ihme-data/united-states-physical-activity-and-obesity-prevalence-county-2001-2011.
IHME (2013b). United States Hypertension Estimates by County 2001–2009. Available at http://ghdx.healthdata.org/record/ihme-data/united-states-hypertension-estimates-county-2001-2009.

IHME (2014). United States Smoking Prevalence by County 1996–2012. Available at http://ghdx.healthdata.org/record/ihme-data/united-states-smoking-prevalence-county-1996-2012.
IHME (2015). United States Alcohol Use Prevalence by County 2002–2012. Available at http://ghdx.healthdata.org/record/ihme-data/united-states-alcohol-use-prevalence-county-2002-2012.
IHME (2016). Diagnosed and Undiagnosed Diabetes Prevalence by County in the U.S., 1999–2012. Available at http://ghdx.healthdata.org/record/ihme-data/united-states-diabetes-prevalence-county-1999-2012.
JOHNSON, M., CAO, J. and KANG, H. (2022). Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment. *Ann. Appl. Stat.* **16** 1111–1129. MR4438826 https://doi.org/10.1214/21-aoas1535
KFF (2020). Status of State Action on the Medicaid Expansion Decision. Available at https://www.kff.org/health-reform/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/. Date Accessed: 8/5/2020.
KFF (2021). Who Could Get Covered Under Medicaid Expansion? State Fact Sheets. Available at https://www.kff.org/medicaid/fact-sheet/uninsured-adults-in-states-that-did-not-expand-who-would-become-eligible-for-medicaid-under-expansion.
KFF (2023). Status of State Medicaid Expansion Decisions: Interactive Map. Available at https://www.kff.org/medicaid/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/. Date Accessed: 2/28/2024.
KHATANA, S. A. M., BHATLA, A., NATHAN, A. S., GIRI, J., SHEN, C., KAZI, D. S., YEH, R. W. and GROENEVELD, P. W. (2019). Association of medicaid expansion with cardiovascular mortality. *JAMA Cardiol.* **4** 671–679.
KILCIOGLU, C. and ZUBIZARRETA, J. R. (2016). Maximizing the information content of a balanced matched sample in a study of the economic performance of green buildings. *Ann. Appl. Stat.* **10** 1997–2020. MR3592046 https://doi.org/10.1214/16-AOAS962
LAM, M. B., PHELAN, J., ORAV, E. J., JHA, A. K. and KEATING, N. L. (2020). Medicaid expansion and mortality among patients with breast, lung, and colorectal cancer. *JAMA Netw. Open* **3** e2024366. https://doi.org/10.1001/jamanetworkopen.2020.24366
LAROCQUE, D. (2005). The Wilcoxon signed-rank test for cluster correlated data. In *Statistical Modeling and Analysis for Complex Data Problems* (P. Duchesne and B. RÉMillard, eds.). *GERAD* 25*th Anniv. Ser.* **1** 309–323. Springer, New York. MR2189542 https://doi.org/10.1007/0-387-24555-3_15
LEE, B. P., DODGE, J. L. and TERRAULT, N. A. (2022). Medicaid expansion and variability in mortality in the USA: A national, observational cohort study. *Lancet Public Health* **7** e48–e55.
LINDROOTH, R. C., PERRAILLON, M. C., HARDY, R. Y. and TUNG, G. J. (2018). Understanding the relationship between medicaid expansions and hospital closures. *Health Aff.* **37** 111–120.
LYCURGUS, T., HANSEN, B. B. and WHITE, M. (2022). Conjuring power from a theory of change: The PWRD method for trials with anticipated variation in effects. *J. Res. Educ. Eff.* **0** 1–27.
MANN, C. Z., HANSEN, B. B. and GAYDOSH, L. (2024). Supplement to "Early effects of 2014 U.S. Medicaid expansions on mortality: Design-based inference for impacts on small subgroups despite small-cell suppression." https://doi.org/10.1214/24-AOAS1910SUPPA, https://doi.org/10.1214/24-AOAS1910SUPPB, https://doi.org/10.1214/24-AOAS1910SUPPC https://doi.org/10.1214/24-AOAS1910SUPPD
MANN, C. Z., HANSEN, B. B., GAYDOSH, L. and LYCURGUS, T. (2021). Protocol—evaluating the effect of ACA medicaid expansion on mortality during the COVID-19 pandemic using county-level matching. *Obs. Stud.* **7** S1–S31.
MANN, H. B. and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18** 50–60. MR0022058 https://doi.org/10.1214/aoms/1177730491
MANTEL, N. (1967). Ranking procedures for arbitrarily restricted observation. *Biometrics* **23** 65–78.
MATTHEWS, K. A., GAGLIOTI, A. H., HOLT, J. B., MCGUIRE, L. C. and GREENLUND, K. J. (2021). County-level concentration of selected chronic conditions among medicare fee-for-service beneficiaries and its association with medicare spending in the United States, 2017. *Popul. Health Manag.* **24** 214–221. https://doi.org/10.1089/pop.2019.0231
METZL, J. (2019). *Dying of Whiteness*. Basic Books, New York.
MILLER, S., ALTEKRUSE, S., JOHNSON, N. and WHERRY, L. R. (2019). Medicaid and mortality: New evidence from linked survey and administrative data. Working Paper No. 26081, National Bureau of Economic Research.
MILLER, S., JOHNSON, N. and WHERRY, L. R. (2021). Medicaid and mortality: New evidence from linked survey and administrative data. *Q. J. Econ.* **136** 1783–1829.
MONTEZ, J. K., MEHRI, N., MONNAT, S. M., BECKFIELD, J., CHAPMAN, D., GRUMBACH, J. M., HAYWARD, M. D., WOOLF, S. H. and ZAJACOVA, A. (2022). U.S. state policy contexts and mortality of working-age adults. *PLoS ONE* **17** e0275466. https://doi.org/10.1371/journal.pone.0275466

NATTINO, G., LU, B., SHI, J., LEMESHOW, S. and XIANG, H. (2021). Triplet matching for estimating causal effects with three treatment arms: A comparative study of mortality by trauma center level. *J. Amer. Statist. Assoc.* **116** 44–53. MR4227673 https://doi.org/10.1080/01621459.2020.1737078

POPESCU, I., DUFFY, E., MENDELSOHN, J. and ESCARCE, J. J. (2018). Racial residential segregation, socioeconomic disparities, and the white-black survival gap. *PLoS ONE* **13** e0193222. https://doi.org/10.1371/journal.pone.0193222

PRESS, C. Q. (2020). Voting and Elections Collection. Available at https://library.cqpress.com/elections/. Date Accessed: 6/11/2020.

ROSENBAUM, P. R. (1991a). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

ROSENBAUM, P. R. (1991b). Some poset statistics. *Ann. Statist.* **19** 1091–1097. MR1105865 https://doi.org/10.1214/aos/1176348141

ROSENBAUM, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253. MR1245357

ROSENBAUM, P. R. (1994). Coherence in observational studies. *Biometrics* **50** 368–374.

ROSENBAUM, P. R. (2001). Observational studies: Overview. In *International Encyclopedia of the Social & Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.) 10808–10815. Elsevier, New York.

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

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

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727 https://doi.org/10.1093/biomet/asm085

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

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSSEN, L. M., HEDEGAARD, H., KHAN, D. and WARNER, M. (2018). County-level trends in suicide rates in the U.S., 2005-2015. *Am. J. Prev. Med.* **55** 72–79. https://doi.org/10.1016/j.amepre.2018.03.020

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

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

SEMPRINI, J. and OLOPADE, O. (2020). Evaluating the effect of medicaid expansion on black/white breast cancer mortality disparities: A difference-in-difference analysis. *JCO Glob. Oncol.* **6** 1178–1183. https://doi.org/10.1200/GO.20.00068

SMALL, D. S., TEN HAVE, T. R. and ROSENBAUM, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *J. Amer. Statist. Assoc.* **103** 271–279. MR2420232 https://doi.org/10.1198/016214507000000897

SONI, A., HENDRYX, M. and SIMON, K. (2017). Medicaid expansion under the affordable care act and insurance coverage in rural and urban areas. *J. Rural Health* **33** 217–226.

SURGEON GENERAL'S ADVISORY COMMITTEE (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Public Health Service Publication 1103. US Department of Health, Education, and Welfare.

SWAMINATHAN, S., SOMMERS, B. D., THORSNESS, R., MEHROTRA, R., LEE, Y. and TRIVEDI, A. N. (2018). Association of medicaid expansion with 1-year mortality among patients with end-stage renal disease. *JAMA* **320** 2242–2250.

TIWARI, C., BEYER, K. and RUSHTON, G. (2014). The impact of data suppression on local mortality rates: The case of CDC WONDER. *Amer. J. Publ. Health* **104** 1386–1388. https://doi.org/10.2105/AJPH.2014.301900

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*, 1 ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VENKATARAMANI, A. S. and CHATTERJEE, P. (2019). Early medicaid expansions and drug overdose mortality in the USA: A quasi-experimental analysis. *J. Gen. Intern. Med.* **34** 23–25.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1** 80–83.

WOOLF, S. H. (2023). Falling behind: The growing gap in life expectancy between the United States and other countries, 1933-2021. *Amer. J. Publ. Health* **113** 970–980. https://doi.org/10.2105/AJPH.2023.307310

YE, W. and RODRIGUEZ, J. M. (2021). Highly vulnerable communities and the affordable care act: Health insurance coverage effects, 2010–2018. *Soc. Sci. Med.* **270** 113670.

YU, R. (2021). Evaluating and improving a matched comparison of antidepressants and bone density. *Biometrics* **77** 1276–1288. MR4357837 https://doi.org/10.1111/biom.13374

YU, R., SMALL, D. S. and ROSENBAUM, P. R. (2021). The information in covariate imbalance in studies of hormone replacement therapy. *Ann. Appl. Stat.* **15** 2023–2042. MR4355087 https://doi.org/10.1214/21-aoas1448

ZHANG, K., TRASKIN, M. and SMALL, D. S. (2012). A powerful and robust test statistic for randomization inference in group-randomized trials with matched pairs of groups. *Biometrics* **68** 75–84. MR2909855 https://doi.org/10.1111/j.1541-0420.2011.01622.x

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2909–2927
https://doi.org/10.1214/24-AOAS1911

MODELS WITH OBSERVATION ERROR AND TEMPORARY EMIGRATION FOR COUNT DATA

BY FABIAN R. KETWAROO[1,a], ELENI MATECHOU[2,b], REBECCA BIDDLE[3,c], SIMON TOLLINGTON[4,d] AND MARIA L. DA SILVA[5,e]

[1]*Swiss Ornithological Institute, Sempach, Switzerland,* [a]*fabian.ketwaroo@vogelwarte.ch*
[2]*School of Mathematics, Statistics and Actuarial Science, University of Kent,* [b]*e.matechou@kent.ac.uk*
[3]*Twycross Zoo,* [c]*rebecca.biddle@twycrosszoo.org*
[4]*School of Animal, Rural and Environmental Sciences, Nottingham Trent University,* [d]*simon.tollington@ntu.ac.uk*
[5]*Laboratory of Ornithology and Bioacoustics, Institute of Biological Sciences, Federal University of Pará,* [e]*mluisa@ufpa.br*

Count data at surveyed sites are an important monitoring tool for several species around the world. However, the raw count data are an underestimate of the size of the monitored population at any one time, as individuals can temporarily leave the site (temporary emigration, TE) and because the probability of detection of individuals, even when using the site, is typically much lower than one (observation error). In this paper we develop a novel modelling framework for estimating population size, from count data, while accounting for both TE and observation error. Our framework builds on the popular class of N-mixture models but extends them in a number of ways. Specifically, we introduce two model classes for TE, a parametric, which relies on temporal models, and a nonparametric, which relies on Dirichlet process mixture models. Both model classes give rise to interesting ecological interpretations of the TE pattern while being parsimonious in terms of the number of parameters required to model the pattern. When accounting for observation error, we use mixed-effects models and implement an efficient Bayesian variable selection algorithm for identifying important predictors for the probability of detection. We demonstrate our new modelling framework using an extensive simulation study, which highlights the importance of using mixed-effects models for the probability of detection and illustrates the performance of the model when estimating population size and underlying TE patterns. We also assess the ability of the corresponding variable selection algorithm to identify important predictors under different scenarios for observation error and its corresponding model. When fitted to two motivating data sets of parrots counted at their roosts, our results provide new insights into how each species uses the roost throughout the year, on changes in population size between and within years, and on observation error.

REFERENCES

ALMOND, R. E., GROOTEN, M. and PETERSON, T. (2020). *Living Planet Report* 2020—*Bending the Curve of Biodiversity Loss*. World Wildlife Fund, Gland, Switzerland.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BEAVER, J. T., BALDWIN, R. W., MESSINGER, M., NEWBOLT, C. H., DITCHKOFF, S. S. and SILMAN, M. R. (2020). Evaluating the use of drones equipped with thermal sensors as an effective method for estimating wildlife. *Wildl. Soc. Bull.* **44** 434–443.

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

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–20. MR1105822 https://doi.org/10.1007/BF00116466

Key words and phrases. N-mixture model, roost counts, Bayesian variable selection, Dirichlet process, temporal models, population size.

BIDDLE, R., PONCE, I. S., CUN, P., TOLLINGTON, S., JONES, M., MARSDEN, S., DEVENISH, C., HORSTMAN, E., BERG, K. et al. (2020). Conservation status of the recently described Ecuadorian Amazon parrot Amazona lilacina. *Bird Conserv. Int.* **30** 586–598.

BIDDLE, R., SOLIS-PONCE, I., JONES, M., MARSDEN, S., PILGRIM, M. and DEVENISH, C. (2021a). The value of local community knowledge in species distribution modelling for a threatened Neotropical parrot. *Biodivers. Conserv.* **30** 1803–1823.

BIDDLE, R., SOLIS-PONCE, I., JONES, M., PILGRIM, M. and MARSDEN, S. (2021b). Parrot ownership and capture in coastal Ecuador: Developing a trapping pressure index. *Diversity* **13** 15.

CARDINALE, B. J., DUFFY, J. E., GONZALEZ, A., HOOPER, D. U., PERRINGS, C., VENAIL, P., NARWANI, A., MACE, G. M., TILMAN, D. et al. (2012). Biodiversity loss and its impact on humanity. *Nature* **486** 59–67.

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**.

CASELLA, G., GHOSH, M., GILL, J. and KYUNG, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. MR2719657 https://doi.org/10.1214/10-BA607

CHANDLER, R. B., ROYLE, J. A. and KING, D. I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology* **92** 1429–1435. https://doi.org/10.1890/10-2433.1

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

DAIL, D. and MADSEN, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics* **67** 577–587. MR2829026 https://doi.org/10.1111/j.1541-0420.2010.01465.x

DE MOURA, L. N., VIELLIARD, J. M. and DA SILVA, M. L. (2010). Seasonal fluctuation of the orange-winged Amazon at a roosting site in Amazonia. *Wilson J. Ornithol.* **122** 88–94.

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

DÉNES, F. V., TELLA, J. L. and BEISSINGER, S. R. (2018). Revisiting methods for estimating parrot abundance and population size. *Emu* **118** 67–79.

DORAZIO, R. M. (2009). On selecting a prior for the precision parameter of Dirchlet process mixture models. *J. Statist. Plann. Inference* **139** 3384–3390. MR2538090 https://doi.org/10.1016/j.jspi.2009.03.009

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FAHRMEIR, L. and LANG, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 201–220. MR1833273 https://doi.org/10.1111/1467-9876.00229

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

FRÜHWIRTH-SCHNATTER, S. and MALSINER-WALLI, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv. Data Anal. Classif.* **13** 33–64. MR3935190 https://doi.org/10.1007/s11634-018-0329-y

GILBERT, N. A., CLARE, J. D. J., STENGLEIN, J. L. and ZUCKERBERG, B. (2021). Abundance estimation of unmarked animals based on camera-trap data. *Conserv. Biol.* **35** 88–100. https://doi.org/10.1111/cobi.13517

GRIFFIN, J. E., MATECHOU, E., BUXTON, A. S., BORMPOUDAKIS, D. and GRIFFITHS, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 377–392. MR4098953 https://doi.org/10.1111/rssc.12390

HAINING, R. P. and LI, G. (2020). *Regression Modelling Wih Spatial and Spatial–Temporal Data*: *A Bayesian Approach*. CRC Press, Boca Raton.

JACKSON, R. M., ROE, J. D., WANGCHUK, R. and HUNTER, D. O. (2006). Estimating snow leopard population abundance using photography and capture–recapture techniques. *Wildl. Soc. Bull.* **34** 772–781.

JETZ, W., MCGEOCH, M. A., GURALNICK, R., FERRIER, S., BECK, J., COSTELLO, M. J., FERNANDEZ, M., GELLER, G. N., KEIL, P. et al. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nat. Ecol. Evol.* **3** 539–551.

KARANTH, K. U. (1995). Estimating tiger Panthera tigris populations from camera-trap data using capture–recapture models. *Biol. Conserv.* **71** 333–338.

KÉRY, M. and ROYLE, J. A. (2016). *Applied Hierarchical Modeling in Ecology—Analysis of Distribution, Abundance and Species Richness in R and BUGS. Vol.* 1: *Prelude and Static Models*. Elsevier/Academic Press, London. With a foreword by Richard Chandler. MR3616659

KÉRY, M. and ROYLE, J. A. (2020). *Applied Hierarchical Modeling in Ecology*: *Analysis of Distribution, Abundance and Species Richness in R and BUGS. Volume* 2: *Dynamic and Advanced Models*. Academic Press, San Diego.

KETWAROO, F. R. MATECHOU, E., BIDDLE, R. TOLLINGTON, S. and DA SILVA, M. L. (2024). Supplement to "Models with observation error and temporary emigration for count data." https://doi.org/10.1214/24-AOAS1911SUPPA, https://doi.org/10.1214/24-AOAS1911SUPPB

KOH, L. P. and WICH, S. A. (2012). Dawn of drone ecology: Low-cost autonomous aerial vehicles for conservation. *Trop. Conserv. Sci.* **5** 121–132.

KOTTAS, A. (2006). Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation. In *Workshop on Learning with Nonparametric Bayesian Methods*, 23*rd International Conference on Machine Learning* (*ICML*) **47**.

LINK, W. A., SCHOFIELD, M. R., BARKER, R. J. and SAUER, J. R. (2018). On the robustness of N-mixture models. *Ecology* **99** 1547–1551. https://doi.org/10.1002/ecy.2362

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

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1032. With comments by James Berger and C. L. Mallows and with a reply by the authors. MR0997578

NAKASHIMA, Y. (2020). Potentiality and limitations of N-mixture and Royle–Nichols models to estimate animal abundance based on noninstantaneous point surveys. *Popul. Ecol.* **62** 151–157.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

NEUBAUER, G., WOLSKA, A., ROWIŃSKI, P. and WESOŁOWSKI, T. (2022). N-mixture models estimate abundance reliably: A field test on Marsh Tit using time-for-space substitution. *Condor* **124** duab054.

PACIOREK, C. (2009). Technical Vignette 5: Understanding intrinsic Gaussian Markov random field spatial models, including intrinsic conditional autoregressive models. Technical report.

POLLOCK, K. H. (1982). A capture–recapture design robust to unequal probability of capture. *J. Wildl. Manag.* **46** 752–757.

ROSS, S. R.-J., O'CONNELL, D. P., DEICHMANN, J. L., DESJONQUÈRES, C., GASC, A., PHILLIPS, J. N., SETHI, S. S., WOOD, C. M. and BURIVALOVA, Z. (2023). Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Funct. Ecol.* **37** 959–975.

ROYLE, J. A. (2004). *N*-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625 https://doi.org/10.1111/j.0006-341X.2004.00142.x

SCHMELLER, D. S., HENRY, P.-Y., JULLIARD, R., GRUBER, B., CLOBERT, J., DZIOCK, F., LENGYEL, S., NOWICKI, P., DERI, E. et al. (2009). Advantages of volunteer-based biodiversity monitoring in Europe. *Conserv. Biol.* **23** 307–316.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems* **17**.

THOMAS, C. D. (2013). Local diversity stays about the same, regional diversity increases, and global diversity declines. *Proc. Natl. Acad. Sci. USA* **110** 19187–19188.

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

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

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

WEST, M. and ESCOBAR, M. D. (1993). Hierarchical priors and mixture models, with application in regression and density estimation. Institute of Statistics and Decision Sciences, Duke Univ.

WILLIAMS, B. K., NICHOLS, J. D. and CONROY, M. J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego.

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

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZWERTS, J. A., STEPHENSON, P., MAISELS, F., ROWCLIFFE, M., ASTARAS, C., JANSEN, P. A., VAN DER WAARDE, J., STERCK, L. E., VERWEIJ, P. A. et al. (2021). Methods for wildlife monitoring in tropical forests: Comparing human observations, camera traps, and passive acoustic sensors. *Conserv. Sci. Pract.* **3** e568.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2928–2949
https://doi.org/10.1214/24-AOAS1915

MULTISITE DISEASE ANALYTICS WITH APPLICATIONS TO ESTIMATING COVID-19 UNDETECTED CASES IN CANADA

BY MATTHEW R. P. PARKER[1,a], JIGUO CAO[1,b], LAURA L. E. COWEN[2,d], LLOYD T. ELLIOTT[1,c] AND JUNLING MA[2,e]

[1]*Department of Statistics and Actuarial Sciences, Simon Fraser University,* [a]*mrparker909@gmail.com,* [b]*jiguo_cao@sfu.ca,* [c]*lloyde@sfu.ca*

[2]*Department of Mathematics and Statistics, University of Victoria,* [d]*lcowen@uvic.ca,* [e]*junlingm@uvic.ca*

Even with daily case counts, the true scope of the COVID-19 pandemic in Canada is unknown due to undetected cases. We develop a novel multisite disease analytics model which estimates undetected cases using discrete-valued multivariate time series in the framework of Bayesian hidden Markov modelling techniques. We apply our multisite model to estimate the pandemic scope using publicly available disease count data including detected cases, recoveries among detected cases, and total deaths. These counts are used to estimate the case detection probability, the infection fatality rate through time, the probability of recovery, and several important population parameters including the rate of spread and importation of external cases. We estimate the total number of active COVID-19 cases per region of Canada for each reporting interval. We applied this multisite model Canada-wide to all provinces and territories, providing an estimate of the total COVID-19 burden for the 90 weeks from 23 April 2020 to 10 February 2022. We also applied this model to the five health authority regions of British Columbia, Canada, describing the pandemic in B.C. over the 31 weeks from 2 April 2020 to 30 October 2020.

REFERENCES

ABDOLLAHI, E., CHAMPREDON, D., LANGLEY, J. M., GALVANI, A. P. and MOGHADAS, S. M. (2020). Temporal estimates of case-fatality rate for COVID-19 outbreaks in Canada and the United States. *CMAJ, Can. Med. Assoc. J.* **192** E666–E670. https://doi.org/10.1503/cmaj.200711

ALENE, M., YISMAW, L., ASSEMIE, M. A., KETEMA, D. B., MENGIST, B., KASSIE, B. and BIRHAN, T. Y. (2021). Magnitude of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLoS ONE* **16** e0249090. https://doi.org/10.1371/journal.pone.0249090

APPLEBY, J. A., KING, N., SAUNDERS, K. E., BAST, A., RIVERA, D., BYUN, J., CUNNINGHAM, S., KHERA, C. and DUFFY, A. C. (2022). Impact of the COVID-19 pandemic on the experience and mental health of university students studying in Canada and the UK: A cross-sectional study. *BMJ Open* **12** e050187.

ARAF, Y., AKTER, F., TANG, Y.-D., FATEMI, R., PARVEZ, M. S. A., ZHENG, C. and HOSSAIN, M. G. (2022). Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. *J. Med. Virol.* **94** 1825–1832.

BC CENTRE FOR DISEASE CONTROL (2020). BC COVID-19 data [surveillance reports]. Retrieved from http://www.bccdc.ca/health-info/diseases-conditions/covid-19/data. Accessed: 2022-02-25.

BÉLAND, D., DINAN, S., ROCCO, P. and WADDAN, A. (2021). Social policy responses to COVID-19 in Canada and the United States: Explaining policy variations between two liberal welfare state regimes. *Soc. Policy Adm.* **55** 280–294.

BENDAVID, E., MULANEY, B., SOOD, N., SHAH, S., BROMLEY-DULFANO, R., LAI, C. et al. (2021). COVID-19 antibody seroprevalence in Santa Clara County, California. *Int. J. Epidemiol.* **50** 410–419.

BUITRAGO-GARCIA, D., EGLI-GANY, D., COUNOTTE, M. J., HOSSMANN, S., IMERI, H., IPEKCI, A. M., SALANTI, G. and LOW, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Med.* **17** e1003346.

CHIMMULA, V. K. R. and ZHANG, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **135** 109864.

Key words and phrases. Infectious disease modelling, hidden Markov models, disease parameter estimation, infection fatality rate.

CHISALE, M. R. O., RAMAZANU, S., MWALE, S. E., KUMWENDA, P., CHIPETA, M., KAMINGA, A. C., NKHATA, O., NYAMBALO, B., CHAVURA, E. et al. (2022). Seroprevalence of anti-SARS-CoV-2 antibodies in Africa: A systematic review and meta-analysis. *Rev. Med. Virol.* **32** e2271.

CYPRESS, B. S. (2022). COVID-19: The economic impact of a pandemic on the healthcare delivery system in the United States. *Nurs. Forum* **57** 323–327. https://doi.org/10.1111/nuf.12677

DE VALPINE, P., PACIOREK, C., TUREK, D., MICHAUD, N., ANDERSON-BERGMAN, C., OBERMEYER, F., WEHRHAHN CORTES, C., RODRÌGUEZ, A., TEMPLE LANG, D. et al. (2021). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. R package version 0.11.1.

DE VALPINE, P., TUREK, D., PACIOREK, C., 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.

DESSON, Z., WELLER, E., MCMEEKIN, P. and AMMI, M. (2020). An analysis of the policy responses to the COVID-19 pandemic in France, Belgium, and Canada. *Health Public Technol.* **9** 430–446.

DIRENZO, G. V., CHE-CASTALDO, C., SAUNDERS, S. P., GRANT, E. H. C. and ZIPKIN, E. F. (2019). Disease-structured N-mixture models: A practical guide to model disease dynamics using count data. *Ecol. Evol.* **9** 899–909.

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.

DOUGHERTY, B. P., SMITH, B. A., CARSON, C. A. and OGDEN, N. H. (2021). Exploring the percentage of COVID-19 cases reported in the community in Canada and associated case fatality ratios. *Infect. Dis. Model.* **6** 123–132. https://doi.org/10.1016/j.idm.2020.11.008

DOZOIS, D. J. A. (2021). Anxiety and depression in Canada during the COVID-19 pandemic: A national survey. *Can. Psychol.* **62** 136–142.

FEIKIN, D. R., WIDDOWSON, M.-A. and MULHOLLAND, K. (2020). Estimating the percentage of a population infected with SARS-CoV-2 using the number of reported deaths: A policy planning tool. *Pathogens* **9** 838.

FERNÁNDEZ-FONTELO, A., CABAÑA, A., PUIG, P. and MORIÑA, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Stat. Med.* **35** 4875–4890. https://doi.org/10.1002/sim.7026

FERNÁNDEZ-FONTELO, A., MORIÑA, D., CABAÑA, A., ARRATIA, A. and PUIG, P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE* **15** e0242956.

FISMAN, D. N., DREWS, S. J., TUITE, A. R. and O'BRIEN, S. F. (2020). Age-specific SARS-CoV-2 infection fatality and case identification fraction in Ontario, Canada Technical Report medRxiv.

GELMAN, A., HWANG, J. and VEHTARI, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24** 997–1016.

GOVERNMENT OF BRITISH COLUMBIA (2020). Phase 1: BC's restart plan. Retrieved from https://www2.gov.bc.ca/gov/content/safety/emergency-preparedness-response-recovery/covid-19-provincial-support/phase-1. Accessed: 2020-12-08.

GOVERNMENT OF CANADA (2022a). COVID-19 daily epidemiology update. Web Archive: /web/20220114205328/ https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html?redir=1. Accessed: 2022-01-14.

GOVERNMENT OF CANADA (2022b). COVID-19 vaccination in Canada. Web Archive: /web/20220120212246/ https://health-infobase.canada.ca/covid-19/vaccination-coverage/. Accessed 2022-01-20.

GOVERNMENT OF YUKON (2022). COVID-19 data dashboard. Web: https://covid-19-data-dashboard.service.yukon.ca/. Accessed 2022-02-18.

HALILI, R., BUNJAKU, J., GASHI, B., HOXHA, T., KAMBERI, A., HOTI, N., AGAHI, R., BASHA, V., BERISHA, V. et al. (2022). Seroprevalence of anti-SARS-CoV-2 antibodies among staff at primary healthcare institutions in Prishtina. *BMC Infect. Dis.* **22** 57.

HASAN, T., PHAM, T. N., NGUYEN, T. A., HIEN THI, T. L., DUYET, V. L., THUY, T. D. et al. (2021). Seroprevalence of SARS-CoV-2 antibodies in high-risk populations in Vietnam. *Int. J. Environ. Res. Public Health* **18** 6353.

HE, J., GUO, Y., MAO, R. and ZHANG, J. (2021). Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. *J. Med. Virol.* **93** 820–830.

HUO, X., CHEN, J. and RUAN, S. (2021). Estimating asymptomatic, undetected and total cases for the COVID-19 outbreak in Wuhan: A mathematical modeling study. *BMC Infect. Dis.* **21** 476.

IOANNIDIS, J. P. A., AXFORS, C. and CONTOPOULOS-IOANNIDIS, D. G. (2020). Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *Environ. Res.* **188** 109890.

KAHN, F., BONANDER, C., MOGHADDASSI, M., RASMUSSEN, M., MALMQVIST, U., INGHAMMAR, M. and BJÖRK, J. (2022). Risk of severe COVID-19 from the Delta and Omicron variants in relation to vaccination status, sex, age and comorbidities - surveillance results from southern Sweden, July 2021 to January 2022. *Euro Surveill.* **27**. https://doi.org/10.2807/1560-7917.ES.2022.27.9.2200121

LAURING, A. S., TENFORDE, M. W., CHAPPELL, J. D., GAGLANI, M., GINDE, A. A. and SELF, W. H. (2022). Clinical severity of, and effectiveness of mRNA vaccines against, covid-19 from omicron, delta, and alpha SARS-CoV-2 variants in the United States: Prospective observational study. *BMJ* **376** e069761.

LI, C., ZHU, Y., QI, C., LIU, L., ZHANG, D., WANG, X., SHE, K., JIA, Y., LIU, T. et al. (2021). Estimating the prevalence of asymptomatic COVID-19 cases and their contribution in transmission—using Henan province, China, as an example. *Front. Med.* **8**.

MAHUMUD, R. A., ALI, M. A., KUNDU, S., RAHMAN, M. A., KAMARA, J. K. and RENZAHO, A. M. N. (2022). Effectiveness of COVID-19 vaccines against delta variant (B.1.617.2): A meta-analysis. *Vaccines* **10** 277.

MORIÑA, D., FERNÁNDEZ-FONTELO, A., CABAÑA, A., ARRATIA, A., ÁVALOS, G. and PUIG, P. (2021). Cumulated burden of Covid-19 in Spain from a Bayesian perspective. *Eur. J. Public Health* **31** 917–920.

MULLAH, M. A. S. and YAN, P. (2022). A semi-parametric mixed model for short-term projection of daily COVID-19 incidence in Canada. *Epidemics* **38** 100537. https://doi.org/10.1016/j.epidem.2022.100537

NATIONAL COLLABORATING CENTRE FOR INFECTIOUS DISEASES (2022). Updates on COVID-19 Variants of Concern (VOC). Retrieved from https://nccid.ca/covid-19-variants/. Accessed: 2022-03-09.

PARKER, M. R. P., CAO, J., COWEN, L. L. E., ELLIOTT, L. T. and MA, J. (2024a). Code supplement to "Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada." https://doi.org/10.1214/24-AOAS1915SUPPA

PARKER, M. R. P., CAO, J., COWEN, L. L. E., ELLIOTT, L. T. and MA, J. (2024b). Appendix supplement to "Multi-site disease analytics with applications to estimating COVID-19 undetected cases in Canada." https://doi.org/10.1214/24-AOAS1915SUPPB

PARKER, M. R. P., LI, Y., ELLIOTT, L. T., MA, J. and COWEN, L. L. E. (2021). Under-reporting of COVID-19 in the northern health authority region of British Columbia. *Canad. J. Statist.* **49** 1018–1038.

PROVINCE OF BRITISH COLUMBIA (2020). BC COVID-19—laboratory information. Retrieved from https://governmentofbc.maps.arcgis.com/home/item.html?id=ba047e4a9bd24beb9ca6e94c05eddef9. Accessed: 2021-04-19.

R CORE TEAM (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

ROYLE, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. https://doi.org/10.1111/j.0006-341X.2004.00142.x

SAEED, S., DREWS, S. J., PAMBRUN, C., YI, Q.-L., OSMOND, L. and O'BRIEN, S. F. (2021). SARS-CoV-2 seroprevalence among blood donors after the first COVID-19 wave in Canada. *Transfusion* **61** 862–872. https://doi.org/10.1111/trf.16296

SHIM, E. (2021). Regional variability in COVID-19 case fatality rate in Canada, February-December 2020. *Int. J. Environ. Res. Public Health* **18**. https://doi.org/10.3390/ijerph18041839

SUBRAMANIAN, R., HE, Q. and PASCUAL, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proc. Natl. Acad. Sci. USA* **118** e2019716118.

TANAKA, S. (2022). Economic impacts of SARS/MERS/COVID-19 in Asian countries. *Asian Econ. Policy Rev.* **17** 41–61.

TUITE, A. R., FISMAN, D. N. and GREER, A. L. (2020). Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ, Can. Med. Assoc. J.* **192** E497–E505. https://doi.org/10.1503/cmaj.200476

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.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 9.

WU, P., LIU, F., CHANG, Z., LIN, Y., REN, M., ZHENG, C., LI, Y., PENG, Z., QIN, Y. et al. (2021). Assessing asymptomatic, presymptomatic, and symptomatic transmission risk of severe acute respiratory syndrome coronavirus 2. *Clin. Infect. Dis.* **73** e1314–e1320.

ZINSZER, K., MCKINNON, B., BOURQUE, N., PIERCE, L., SAUCIER, A., OTIS, A., CHERIET, I., PAPENBURG, J., HAMELIN, M. et al. (2021). Seroprevalence of SARS-CoV-2 antibodies among children in school and day care in Montreal, Canada. *JAMA Netw. Open* **4** e2135975. https://doi.org/10.1001/jamanetworkopen.2021.35975

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2950–2978
https://doi.org/10.1214/24-AOAS1916

BACKGROUND MODELING FOR DOUBLE HIGGS BOSON PRODUCTION: DENSITY RATIOS AND OPTIMAL TRANSPORT

BY TUDOR MANOLE[1,a], PATRICK BRYANT[2,b], JOHN ALISON[2,c], MIKAEL KUUSELA[3,d] AND LARRY WASSERMAN[3,e]

[1]*Statistics and Data Science Center, Massachusetts Institute of Technology,* [a]*tmanole@mit.edu*

[2]*Department of Physics and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University,* [b]*pbryant2@andrew.cmu.edu,* [c]*johnalison@cmu.edu*

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

We study the problem of data-driven background estimation, arising in the search of physics signals predicted by the Standard Model at the Large Hadron Collider. Our work is motivated by the search for the production of pairs of Higgs bosons decaying into four bottom quarks. A number of other physical processes, known as background, also share the same final state. The data arising in this problem is, therefore, a mixture of unlabeled background and signal events, and the primary aim of the analysis is to determine whether the proportion of unlabeled signal events is nonzero. A challenging but necessary first step is to estimate the distribution of background events. Past work in this area has determined regions of the space of collider events, where signal is unlikely to appear and where the background distribution is, therefore, identifiable. The background distribution can be estimated in these regions and extrapolated into the region of primary interest using transfer learning with a multivariate classifier. We build upon this existing approach in two ways. First, we revisit this method by developing a customized residual neural network which is tailored to the structure and symmetries of collider data. Second, we develop a new method for background estimation, based on the optimal transport problem, which relies on modeling assumptions distinct from earlier work. These two methods can serve as cross-checks for each other in particle physics analyses, due to the complementarity of their underlying assumptions. We compare their performance on simulated double Higgs boson data.

REFERENCES

ALISON, J. (2015). The road to discovery: Detector alignment, electron identification, particle misidentification, WW physics, and the discovery of the Higgs boson Ph.D. thesis. Presented 08 Nov 2012.

ALWALL, J., HERQUET, M., MALTONI, F., MATTELAER, O. and STELZER, T. (2011). MadGraph 5: Going beyond. *J. High Energy Phys.* **2011** 128.

ATLAS (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 (2015). Observation and measurement of Higgs boson decays to WW* with the ATLAS detector. *Phys. Rev. D* **92** 012006.

ATLAS (2018a). Search for pair production of higgsinos in final states with at least three b-tagged jets in 13 TeV pp collisions using the ATLAS detector. *Phys. Rev. D* **98**.

ATLAS (2018b). Observation of $H \to bb$ decays and VH production with the ATLAS detector. *Phys. Lett. B* **786** 59–86.

ATLAS (2018c). Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D* **98** 052005.

ATLAS (2018d). Measurements of the Higgs boson production, fiducial and differential cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-018.

Key words and phrases. High energy physics, Large Hadron Collider, optimal transport map, Wasserstein distance, domain adaptation, transfer learning, residual neural network.

ATLAS (2018e). Measurement of gluon fusion and vector-boson-fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. ATLAS-CONF-2018-004.

ATLAS (2019). Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *J. High Energy Phys.* **2019** 30.

ATLAS (2019b). Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D* **99** 072001.

ATLAS (2021). Search for the HH$\to b\bar{b}b\bar{b}$ process via vector-boson fusion production using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *J. High Energy Phys.* **2021**.

ATLAS (2022). Search for non-resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector Technical Report CERN Geneva. ATLAS-CONF-2022-035.

ATLAS, CMS and HIGGS COMBINATION GROUP (2011). Procedure for the LHC Higgs boson search combination in Summer 2011. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11.

BARLOW, R. (1987). Event classification using weighting methods. *J. Comput. Phys.* **72** 202–219.

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.

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method*. *Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BORISYAK, M. and KAZEEV, N. (2019). Machine learning on data with sPlot background subtraction. *J. Instrum.* **14** P08020–P08020.

BREHMER, J., LOUPPE, G., PAVEZ, J. and CRANMER, K. (2020). Mining gold from implicit models to improve likelihood-free inference. *Proc. Natl. Acad. Sci. USA* **117** 5242–5249. MR4225023 https://doi.org/10.1073/pnas.1915980117

BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** 375–417. MR1100809 https://doi.org/10.1002/cpa.3160440402

BRYANT, P. E. (2018). Search for pair production of Higgs bosons in the four bottom quark final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Ph.D. thesis The Univ. Chicago, Chicago, IL.

CACCIARI, M., SALAM, G. P. and SOYEZ, G. (2008). The anti-kt jet clustering algorithm. *J. High Energy Phys.* **2008** 063.

CAI, T., CHENG, J., CRAIG, N. and CRAIG, K. (2020). Linearized optimal transport for collider events. *Phys. Rev. D* **102** 116019.

CHENG, K. F. and CHU, C. K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli* **10** 583–604. MR2076064 https://doi.org/10.3150/bj/1093265631

CHOI, S. and OH, H. (2021). Improved extrapolation methods of data-driven background estimations in high energy physics. *Eur. Phys. J. C* **81** 643.

CMS (2008). The CMS experiment at the CERN LHC. *J. Instrum.* **3** S08004–S08004.

CMS (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 (2017). Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV.

CMS (2018a). Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett.* **121** 121801.

CMS (2018b). Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *J. High Energy Phys.* **2018** 185.

CMS (2018c). Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV. CMS PAS HIG-18-001.

CMS (2018d). Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett. B* **779** 283–316.

CMS (2018e). Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *J. Instrum.* **13** P05011.

CMS (2019). Measurements of properties of the Higgs boson decaying to a W boson pair in pp collisions at $\sqrt{s} =$ 13 TeV. *Phys. Lett. B* **791** 96–129.

CMS (2022). Search for Higgs boson pair production in the four b quark final state in proton-proton collisions at $\sqrt{s} = 13$ TeV.

COURTY, N., FLAMARY, R., TUIA, D. and RAKOTOMAMONJY, A. (2016). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 1853–1865.

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

DE LARA, L., GONZÁLEZ-SANZ, A. and LOUBES, J.-M. (2021). A consistent extension of discrete optimal transport maps for machine learning applications. ArXiv Preprint. Available at arXiv:2102.08644.

DEB, N., GHOSAL, P. and SEN, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Adv. Neural Inf. Process. Syst.* **34**.

DEMBINSKI, H., KENZIE, M., LANGENBRUCH, C. and SCHMELLING, M. (2022). Custom orthogonal weight functions (COWs) for event classification. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* 167270.

DI MICCO, B., GOUZEVITCH, M., MAZZITELLI, J. and VERNIERI, C. (2020). Higgs boson potential at colliders: Status and perspectives. *Rev. Phys.* **5** 100045.

ENGLERT, F. and BROUT, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.* **13** 321–323. MR0174314 https://doi.org/10.1103/PhysRevLett.13.321

FIGALLI, A. (2010). The optimal partial transport problem. *Arch. Ration. Mech. Anal.* **195** 533–560. MR2592287 https://doi.org/10.1007/s00205-008-0212-7

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. Technical Report No. 4, Project No. 21-49-004 USAF School of Aviation Medicine Randolph Field, TX.

FLAMARY, R., COURTY, N., GRAMFORT, A., ALAYA, M. Z., BOISBUNON, A., CHAMBON, S., CHAPEL, L., CORENFLOS, A., FATRAS, K. et al. (2021). POT: Python optimal transport. *J. Mach. Learn. Res.* **22** 1–8.

FORROW, A., HÜTTER, J.-C., NITZAN, M., RIGOLLET, P., SCHIEBINGER, G. and WEED, J. (2019). Statistical optimal transport via factored couplings. In *Proceedings of the* 22*nd International Conference on Artificial Intelligence and Statistics* 2454–2465.

GHOSAL, P. and SEN, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing. *Ann. Statist.* **50** 1012–1037. MR4404927 https://doi.org/10.1214/21-aos2136

GOLDFELD, Z., KATO, K., RIOUX, G. and SADHU, R. (2024). Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electron. J. Stat.* **18** 980–1041. MR4718466 https://doi.org/10.1214/24-ejs2217

GONZÁLEZ-SANZ, A., LOUBES, J.-M. and NILES-WEED, J. (2022). Weak limits of entropy regularized optimal transport; potentials, plans and divergences. ArXiv Preprint. Available at arXiv:2207.07427.

GUNSILIUS, F. and XU, Y. (2021). Matching for causal effects via multimarginal optimal transport. ArXiv Preprint. Available at arXiv:2112.04398.

GUNSILIUS, F. F. (2022). On the convergence rate of potentials of Brenier maps. *Econometric Theory* **38** 381–417. MR4407051 https://doi.org/10.1017/S0266466621000037

HAGEDORN, R. (1963). *Relativistic Kinematics*: *A Guide to the Kinematic Problems of High-Energy Physics*. W. A. Benjamin, New York-Amsterdam. MR0158658

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, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HEINRICH, J. and LYONS, L. (2007). Systematic errors. *Annu. Rev. Nucl. Part. Sci.* **57** 145–169.

HIGGS, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* **13** 508–509. MR0175554 https://doi.org/10.1103/PhysRevLett.13.508

HO, N., HUYNH, V., PHUNG, D. and JORDAN, M. (2019). Probabilistic multilevel clustering via composite transportation distance. In *Proceedings of the* 22*nd International Conference on Artificial Intelligence and Statistics* 3149–3157. PMLR.

HO, N., NGUYEN, X., YUROCHKIN, M., BUI, H. H., HUYNH, V. and PHUNG, D. (2017). Multilevel clustering via Wasserstein means. In *Proceedings of the* 34*th International Conference on Machine Learning* 1501–1509. PMLR.

HÜTTER, J.-C. and RIGOLLET, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.* **49** 1166–1194. MR4255123 https://doi.org/10.1214/20-aos1997

HUYNH, V., HO, N., DAM, N., NGUYEN, X., YUROCHKIN, M., BUI, H. and PHUNG, D. (2021). On efficient multilevel clustering via Wasserstein distances. *J. Mach. Learn. Res.* **22** 145. MR4318501

KANTOROVICH, L. V. (1948). On a problem of Monge. In *CR* (*Doklady*) *Acad. Sci. URSS* (*NS*) **3** 225–226.

KANTOROVICH, L. V. (1942). On the translocation of masses. *C. R.* (*Dokl.*) *Acad. Sci. URSS* **37** 199–201. MR0009619

KASIECZKA, G., NACHMAN, B., SCHWARTZ, M. D. and SHIH, D. (2021). Automating the ABCD method with machine learning. *Phys. Rev. D* **103** 035021. MR4225448 https://doi.org/10.1103/physrevd.103.035021

KLATT, M., TAMELING, C. and MUNK, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.* **2** 419–443. MR4105566 https://doi.org/10.1137/19M1278788

KNOTT, M. and SMITH, C. S. (1984). On the optimal mapping of distributions. *J. Optim. Theory Appl.* **43** 39–49. MR0745785 https://doi.org/10.1007/BF00934745

KOLOURI, S., PARK, S. R., THORPE, M., SLEPCEV, D. and ROHDE, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Process. Mag.* **34** 43–59.

KOMISKE, P. T., MASTANDREA, R., METODIEV, E. M., NAIK, P. and THALER, J. (2020). Exploring the space of jets with CMS open data. *Phys. Rev. D* **101** 034009.
KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2019). Metric space of collider events. *Phys. Rev. Lett.* **123** 041801. https://doi.org/10.1103/PhysRevLett.123.041801
KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2020). The hidden geometry of particle collisions. *J. High Energy Phys.* **7** 6. MR4138090 https://doi.org/10.1007/jhep07(2020)006
KOMISKE, P. T., METODIEV, E. M. and THALER, J. (2022). EnergyFlow Python package. https://energyflow.network/.
KPOTUFE, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* 1320–1328. PMLR.
KRAUSE, C. and SHIH, D. (2023a). Fast and accurate simulations of calorimeter showers with normalizing flows. *Phys. Rev. D* **107** 113003.
KRAUSE, C. and SHIH, D. (2023b). Accelerating accurate simulations of calorimeter showers with normalizing flows and probability density distillation. *Phys. Rev. D* **107** 113004.
KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. MR0075510 https://doi.org/10.1002/nav.3800020109
LEE, J., DABAGIA, M., DYER, E. and ROZELL, C. (2019). Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems* 13453–13463.
LIERO, M., MIELKE, A. and SAVARÉ, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Invent. Math.* **211** 969–1117. MR3763404 https://doi.org/10.1007/s00222-017-0759-8
LYONS, L. (1986). *Statistics for Nuclear and Particle Physicists*. Cambridge Univ. Press, Cambridge.
MAKKUVA, A. V., TAGHVAEI, A., OH, S. and LEE, J. D. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning* 6672–6681. PMLR.
MANOLE, T., BALAKRISHNAN, S., NILES-WEED, J. and WASSERMAN, L. (2024). Plugin estimation of smooth optimal transport maps. *Ann. Statist.* **52** 966–998. https://doi.org/10.1214/24-AOS2379
MANOLE, T., BRYANT, P., ALISON, J., KUUSELA, M. and WASSERMAN, L. (2024). Supplement to "Background modeling for double Higgs boson production: Density ratios and optimal transport." https://doi.org/10.1214/24-AOAS1916SUPPA, https://doi.org/10.1214/24-AOAS1916SUPPB
MONGE, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Hist. Acad. Roy. Sci. Paris*.
NATH, J. S. and JAWANPURIA, P. (2020). Statistical optimal transport posed as learning kernel mean embedding. In *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, *NIPS* '20, *Vancouver*, *BC*, *Canada*. Curran Associates, Red Hook, NY, USA.
PANARETOS, V. M. and ZEMEL, Y. (2019). Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **6** 405–431. MR3939527 https://doi.org/10.1146/annurev-statistics-030718-104938
PANARETOS, V. M. and ZEMEL, Y. (2019b). *An Invitation to Statistics in Wasserstein Space*. Springer, Berlin.
PELE, O. and WERMAN, M. (2008). A linear time histogram metric for improved SIFT matching. In *European Conference on Computer Vision* 495–508. Springer, Berlin.
PELEG, S., WERMAN, M. and ROM, H. (1989). A unified approach to the change of resolution: Space and gray-level. *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 739–742.
PERROT, M., COURTY, N., FLAMARY, R. and HABRARD, A. (2016). Mapping estimation for discrete optimal transport. In *Advances in Neural Information Processing Systems* **29** (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4197–4205. Curran Associates, Red Hook.
PEYRÉ, G. and CUTURI, M. (2019). Computational optimal transport. *Found. Trends Mach. Learn.* **11** 355–607.
PIVK, M. and LE DIBERDER, F. (2005). SPlot: A statistical tool to unfold data distributions. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **555** 356–369.
PLACAKYTE, R. (2011). Parton distribution functions. ArXiv Preprint. Available at arXiv:1111.5452.
POLLARD, C. and WINDISCHHOFER, P. (2022). Transport away your problems: Calibrating stochastic simulations with optimal transport. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **1027** 166119.
POOLADIAN, A.-A. and NILES-WEED, J. (2021). Entropic estimation of optimal transport maps. ArXiv Preprint. Available at arXiv:2109.12004.
QIN, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* **85** 619–630. MR1665814 https://doi.org/10.1093/biomet/85.3.619
RAKOTOMAMONJY, A., FLAMARY, R., GASSO, G., EL ALAYA, M., BERAR, M. and COURTY, N. (2022). Optimal transport for conditional domain matching and label shift. *Mach. Learn.* **111** 1651–1670. MR4426353 https://doi.org/10.1007/s10994-021-06088-2
RAMDAS, A., GARCÍA TRILLOS, N. and CUTURI, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** 47. MR3608466 https://doi.org/10.3390/e19020047

READ, A. L. (1999). Linear interpolation of histograms. *Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip.* **425** 357–360.

REDKO, I., HABRARD, A. and SEBBAN, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 737–753. Springer, Berlin.

REISS, R. D. (2012). *A Course on Point Processes*. Springer, Berlin.

RIPPL, T., MUNK, A. and STURM, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. MR3545279 https://doi.org/10.1016/j.jmva.2016.06.005

RUBNER, Y., TOMASI, C. and GUIBAS, L. J. (2000). The Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40** 99–121.

SAKUMA, T. and MCCAULEY, T. (2014). Detector and event visualization with SketchUp at the CMS experiment. In *Journal of Physics*: *Conference Series* **513** 022032. IOP Publishing, Bristol.

SILVERMAN, B. W. and JONES, M. C. (1989). E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). *Int. Stat. Rev.* 233–238.

VILLANI, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483 https://doi.org/10.1090/gsm/058

VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

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

YOSINSKI, J., CLUNE, J., BENGIO, Y. and LIPSON, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* **27**. Curran Associates, Red Hook.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 2979–3006
https://doi.org/10.1214/24-AOAS1917

STATISTICAL CURVE MODELS FOR INFERRING 3D CHROMATIN ARCHITECTURE

BY ELENA TUZHILINA[1,a], TREVOR HASTIE[2,b] AND MARK SEGAL[3,c]

[1]*Department of Statistical Sciences, University of Toronto,* [a]*elena.tuzhilina@utoronto.ca*
[2]*Department of Statistics, Stanford University,* [b]*hastie@stanford.edu*
[3]*Department of Epidemiology and Biostatistics, University of California,* [c]*Mark.Segal@ucsf.edu*

Reconstructing three-dimensional (3D) chromatin structure from conformation capture assays (such as Hi-C) is a critical task in computational biology, since chromatin spatial architecture plays a vital role in numerous cellular processes and direct imaging is challenging. Most existing algorithms that operate on Hi-C contact matrices produce reconstructed 3D configurations in the form of a polygonal chain. However, none of the methods exploit the fact that the target solution is a (smooth) curve in 3D: this contiguity attribute is either ignored or indirectly addressed by imposing spatial constraints that are challenging to formulate. In this paper we develop both B-spline and smoothing spline techniques for directly capturing this potentially complex 1D curve. We subsequently combine these techniques with a Poisson model for contact counts and compare their performance on a real data example. In addition, motivated by the sparsity of Hi-C contact data, especially when obtained from single-cell assays, we appreciably extend the class of distributions used to model contact counts. We build a general distribution-based metric scaling (*DBMS*) framework from which we develop zero-inflated and Hurdle Poisson models as well as negative binomial applications. Illustrative applications make recourse to bulk Hi-C data from IMR90 cells and single-cell Hi-C data from mouse embryonic stem cells.

REFERENCES

AY, F., BUNNIK, E. M., VAROQUAUX, N., BOL, S. M., PRUDHOMME, J., VERT, J. P., NOBLE, W. S. and LE ROCH, K. G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res*. **24** 974–88.

BELYAEVA, A., KUBJAS, K., SUN, L. J. and UHLER, C. (2022). Identifying 3D genome organization in diploid organisms via Euclidean distance geometry. *SIAM J. Math. Data Sci*. **4** 204–228. MR4386482 https://doi.org/10.1137/21M1390372

BENTBIB, A. H. and KANBER, A. (2015). Block power method for SVD decomposition. *An. Ştiinţ. Univ. "Ovidius" Constanţa Ser. Mat*. **23** 45–58. MR3348699 https://doi.org/10.1515/auom-2015-0024

CAPURSO, D., BENGTSSON, H. and SEGAL, M. R. (2016). Discovering hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions. *Nucleic Acids Res*. **44** 2028–2035. https://doi.org/10.1093/nar/gkw070

CAUER, A. G., YARDIMCI, G., VERT, J.-P., VAROQUAUX, N. and NOBLE, W. S. (2019). Inferring diploid 3D chromatin structures from Hi-C data. In 19*th International Workshop on Algorithms in Bioinformatics* (*WABI* 2019) **143** 11:1–11:13.

DUAN, Z., ANDRONESCU, M., SCHUTZ, K., MCILWAIN, S., KIM, Y. J., LEE, C., SHENDURE, J., FIELDS, S., BLAU, C. A. et al. (2010). A three-dimensional model of the yeast genome. *Nature* **465** 363–367. https://doi.org/10.1038/nature08973

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

GURMU, S. (1998). Generalized hurdle count data regression models. *Econom. Lett*. **58** 263–268. https://doi.org/10.1016/S0165-1765(97)00295-4

HASTIE, T. and STUETZLE, W. (1989). Principal curves. *J. Amer. Statist. Assoc*. **84** 502–516. MR1010339

Key words and phrases. Spatial structure, conformation reconstruction, metric scaling, splines.

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

KNIGHT, P. A. and RUIZ, D. (2013). A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33** 1029–1047. MR3081493 https://doi.org/10.1093/imanum/drs019

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

LEE, C. S., WANG, R. W., CHANG, H. H., CAPURSO, D., SEGAL, M. R. and HABER, J. E. (2016). Chromosome position determines the success of double-strand break repair. *Proc. Natl. Acad. Sci.* **113** 146–154.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J. et al. (2009). Comprehensive mapping of long-range contacts reveals folding principles of the human genome. *Science* **326** 289–293.

LUO, H., LI, X., FU, H. and PENG, C. (2020). HiCHap: A package to correct and analyze the diploid Hi-C data. *BMC Genomics* **21** 746.

MARCO, A., MEHARENA, H. S., DILEEP, V., RAJU, R. M., DAVILA-VELDERRAIN, J., ZHANG, A. L., ADAIKKAN, C., YOUNG, J. Z., GAO, F. et al. (2020). Mapping the epigenomic and transcriptomic interplay during memory formation and recall in the hippocampal engram ensemble. *Nat. Neurosci.* **23** 1606–1617. https://doi.org/10.1038/s41593-020-00717-0

OLUWADARE, O., HIGHSMITH, M. and CHENG, J. (2019). An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proced. Online* **21** 1–20.

PARK, J. and LIN, S. (2017). A random effect model for reconstruction of spatial chromatin structure. *Biometrics* **73** 52–62. MR3632351 https://doi.org/10.1111/biom.12544

PAYNE, A. C., CHIANG, Z. D., REGINATO, P. L., MANGIAMELI, S. M., MURRAY, E. M., YAO, C.-C., MARKOULAKI, S., EARL, A. S., LABADE, A. S. et al. (2021). In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**. eaay3446.

RAMANI, V., DENG, X., QIU, R., GUNDERSON, K. L., STEEMERS, F. J., DISTECHE, C. M., NOBLE, W. S., DUAN, Z. and SHENDURE, J. (2017). Massively multiplex single-cell Hi-C. *Nat. Methods* **14** 263–266. https://doi.org/10.1038/nmeth.4155

RAO, S. S. P., HUNTLEY, M. H., DURAND, N. C., STAMENOVA, E. K., BOCHKOV, I. D., ROBINSON, J. T., SANBORN, A. L., MACHOL, I., OMER, A. D. et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159** 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

RIEBER, L. and MAHONY, S. (2017). miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33** 261–266.

ROSENTHAL, M., BRYNER, D., HUFFER, F., EVANS, S., SRIVASTAVA, A. and NERETTI, N. (2019). Bayesian estimation of three-dimensional chromosomal structure from single-cell Hi-C data. *J. Comput. Biol.* **26** 1191–1202. MR4033369 https://doi.org/10.1089/cmb.2019.0100

SEGAL, M. R. (2023). Assessing chromatin relocalization in 3D using the patient rule induction method. *Biostatistics* **24** 618–634. MR4615244 https://doi.org/10.1093/biostatistics/kxab033

KNOPP, P. and SINKHORN, R. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* **21** 343–348. MR0210731

STEVENS, T. J., LANDO, D., BASU, S., ATKINSON, L. P., CAO, Y., LEE, S. F., LEEB, M., WOHLFAHRT, K. J., BOUCHER, W. et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544** 59–64. https://doi.org/10.1038/nature21429

SU, J. H., ZHENG, P., KINROT, S. S., BINTU, B. and ZHUANG, X. (2020). Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* **182** 1641–1659.

TUZHILINA, E. and HASTIE, T. (2021). Weighted Low Rank Matrix Approximation and Acceleration. Available at arXiv:2109.11057.

TUZHILINA, E., HASTIE, T. and SEGAL, M. (2024). Supplement to "Statistical curve models for inferring 3D chromatin architecture." https://doi.org/10.1214/24-AOAS1917SUPPA, https://doi.org/10.1214/24-AOAS1917SUPPB

TUZHILINA, E., HASTIE, T. J. and SEGAL, M. R. (2022). Principal curve approaches for inferring 3D chromatin architecture. *Biostatistics* **23** 626–642. MR4409771 https://doi.org/10.1093/biostatistics/kxaa046

VAROQUAUX, N., AY, F., NOBLE, W. S. and VERT, J. P. (2014). A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30** 26–33.

VAROQUAUX, N., NOBLE, W. S. and VERT, J. P. (2021). Inference of genome 3D architecture by modeling overdispersion of Hi-C data. Available at https://www.biorxiv.org/content/10.1101/2021.02.04.429864v1.

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

YANG, T., ZHANG, F., YARDIMCI, G. G., SONG, F., HARDISON, R. C., NOBLE, W. S., YUE, F. and LI, Q. (2017). HiCRep: Assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res*. **27** 1939–1949.

ZHANG, Z., LI, G., TOH, K.-C. and SUNG, W.-K. (2013). 3D chromosome modeling with semi-definite programming and Hi-C data. *J. Comput. Biol*. **20** 831–846. MR3130294 https://doi.org/10.1089/cmb.2013.0076

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3007–3023
https://doi.org/10.1214/24-AOAS1919

COMMUNICATION NETWORK DYNAMICS IN A LARGE ORGANIZATIONAL HIERARCHY

BY NATHANIEL JOSEPHS[1,a], SIDA PENG[2,b] AND FORREST W. CRAWFORD[3,c]

[1]*Department of Statistics, North Carolina State University,* [a]*nathaniel.josephs@ncsu.edu*
[2]*Office of Chief Economist, Microsoft Research,* [b]*sidpeng@microsoft.com*
[3]*Department of Biostatistics, Yale School of Public Health,* [c]*forrest.crawford@yale.edu*

Most businesses impose a supervisory hierarchy on employees to facilitate management, decision-making, and collaboration, yet routine inter-employee communication patterns within workplaces tend to emerge more naturally as a consequence of both supervisory relationships and the needs of the organization. What then is the relationship between a formal organizational structure and the emergent communications between its employees? Understanding the nature of this relationship is critical for the successful management of an organization. While scholars of organizational management have proposed theories relating organizational trees to communication dynamics, and separately, network scientists have studied the topological structure of communication patterns in different types of organizations; existing empirical analyses are both lacking in representativeness and limited in size. In fact, much of the methodology used to study the relationship between organizational hierarchy and communication patterns (and much of what is known about this relationship) comes from analyses of the Enron email corpus, reflecting a uniquely dysfunctional corporate environment. In this paper we develop new methodology for assessing the relationship between organizational hierarchy and communication dynamics and apply it to Microsoft Corporation, currently the highest valued company in the world, consisting of approximately 200,000 employees divided into 88 teams, organizational trees rooted at the senior leadership level. This reveals distinct communication network structures within and between teams. We then characterize the relationship of routine employee communication patterns to these team supervisory hierarchies, while empirically evaluating several theories of organizational management and performance. To do so, we propose new measures of communication reciprocity and new shortest-path distances for trees to track the frequency of messages passed up, down, and across the organizational hierarchy. By describing how communication clusters around the formal organization, we reveal the emergent communication dynamics between employees and the crucial role of position in the hierarchy.

REFERENCES

ATHREYA, A., LUBBERTS, Z., PARK, Y. and PRIEBE, C. E. (2022). Discovering underlying dynamics in time series of networks. arXiv preprint. Available at arXiv:2205.06877.

BASTIAN, M., HEYMANN, S. and JACOMY, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media* **3** 361–362.

BOEVA, V., LUNDBERG, L., KOTA, S. M. H. and SKÖLD, L. (2017). Analysis of organizational structure through cluster validation techniques: Evaluation of email communications at an organizational level. In 2017 *IEEE International Conference on Data Mining Workshops* (*ICDMW*) 170–176. IEEE, New Orleans, LA, USA. https://doi.org/10.1109/ICDMW.2017.28

BROIDO, A. D. and CLAUSET, A. (2019). Scale-free networks are rare. *Nat. Commun.* **10** 1–10.

Key words and phrases. Communication dynamics, email network, organizational hierarchy, reciprocity, reporting distance, path analysis, latent tree.

CAPOBIANCO, M. F. and MOLLUZZO, J. C. (1979). The strength of a graph and its application to organizational structure. *Soc. Netw.* **2** 275–283.
CLAUSET, A., MOORE, C. and NEWMAN, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature* **453** 98–101. https://doi.org/10.1038/nature06830
CLAUSET, A., SHALIZI, C. R. and NEWMAN, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* **51** 661–703. MR2563829 https://doi.org/10.1137/070710111
COLIZZA, V., FLAMMINI, A., SERRANO, M. A. and VESPIGNANI, A. (2006). Detecting rich-club ordering in complex networks. *Nat. Phys.* **2** 110–115.
CREAMER, G., ROWE, R., HERSHKOP, S. and STOLFO, S. J. (2007). Segmentation and automated social hierarchy detection through email network analysis. In *International Workshop on Social Network Mining and Analysis* 40–58. Springer, Berlin.
CROSS, R., BORGATTI, S. P. and PARKER, A. (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *Calif. Manag. Rev.* **44** 25–46.
CROSS, R., PARISE, S. and WEISS, L. M. (2007). The role of networks in organizational change. *McKinsey Q.* **3** 28–41.
DIESNER, J. and CARLEY, K. M. (2005). Exploration of communication networks from the Enron email corpus. In *SIAM International Conference on Data Mining*: *Workshop on Link Analysis*, *Counterterrorism and Security*, *Newport Beach*, *CA* 3–14. Citeseer, Newport Beach, CA, USA.
DONG, Y., TANG, J., CHAWLA, N. V., LOU, T., YANG, Y. and WANG, B. (2015). Inferring social status and rich club effects in enterprise communication networks. *PLoS ONE* **10** e0119446.
DONNAT, C. and HOLMES, S. (2018). Tracking network dynamics: A survey using graph distances. *Ann. Appl. Stat.* **12** 971–1012.
ECKHAUS, E. and SHEAFFER, Z. (2018). Managerial hubris detection: The case of Enron. *Risk Manag.* **20** 304–325.
EDGE, D., LARSON, J., TRANDEV, N., SHAH, N. P., BURACTAON, C., CAURVINA, N., EVANS, N. and WHITE, C. M. (2020). Workgroup Mapping: Visual Analysis of Collaboration Culture. arXiv preprint. Available at arXiv:2005.00402.
FIRE, M. and PUZIS, R. (2016). Organization mining using online social networks. *Netw. Spat. Econ.* **16** 545–578.
GALBRAITH, J. R. (2008). Organization design. In *Handbook of Organization Development*. 325–352.
GARLASCHELLI, D. and LOFFREDO, M. I. (2004). Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.* **93** 268701.
GILLESPIE, C. S. (2015). Fitting heavy tailed distributions: The poweRlaw package. *J. Stat. Softw.* **64** 1–16. https://doi.org/10.18637/jss.v064.i02
GUIMERA, R., DANON, L., DIAZ-GUILERA, A., GIRALT, F. and ARENAS, A. (2006). The real communication network behind the formal chart: Community structure in organizations. *J. Econ. Behav. Organ.* **61** 653–667.
GUPTE, M., SHANKAR, P., LI, J., MUTHUKRISHNAN, S. and IFTODE, L. (2011). Finding hierarchy in directed online social networks. In *Proceedings of the* 20*th International Conference on World Wide Web* 557–566.
HATCH, M. J. and SCHULTZ, M. (1997). Relations between organizational culture, identity and image. *Eur. J. Mark.*
HOLTZHAUSEN, D. (2002). The effects of a divisionalised and decentralised organisational structure on a formal internal communication function in a South African organisation. *J. Commun. Manag.*
HOSSAIN, L. (2009). Effect of organisational position and network centrality on project coordination. *Int. J. Proj. Manag.* **27** 680–689.
JOSEPHS, N., PENG, S. and CRAWFORD, F. W. (2024). Supplement to "Communication network dynamics in a large organizational hierarchy." https://doi.org/10.1214/24-AOAS1919SUPPA, https://doi.org/10.1214/24-AOAS1919SUPPB
KIM, Y.-W. (1968). Pseudo quasi metric spaces. *Proc. Jpn. Acad.* **44** 1009–1012.
KLIMT, B. and YANG, Y. (2004). The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning* 217–226. Springer, Berlin.
KOLACZYK, E. D. and CSÁRDI, G. (2014). *Statistical Analysis of Network Data with R* **65**. Springer, Berlin.
KRACKHARDT, D. and HANSON, J. R. (1993). Informal networks: The company behind the chart. *Harv. Bus. Rev.* **71** 104–111.
KRACKHARDT, D. and STERN, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Soc. Psychol. Q.* 123–140.
MAIYA, A. S. and BERGER-WOLF, T. Y. (2009). Inferring the maximum likelihood hierarchy in social networks. In 2009 *International Conference on Computational Science and Engineering* **4** 245–250. IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/CSE.2009.235

MICHALSKI, R., PALUS, S. and KAZIENKO, P. (2011). Matching organizational structure and social network extracted from email communication. In *International Conference on Business Information Systems* 197–206. Springer, Berlin.

NAMATA, G., GETOOR, L. and DIEHL, C. (2006). Inferring formal titles in organizational email archives. In *Proc. of the ICML Workshop on Statistical Network Analysis*.

NEWMAN, M. E. (2006). Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103** 8577–8582.

NIELSEN, C. (2016). What work email can reveal about performance and potential. *Harv. Bus. Rev.*

NUREK, M. and MICHALSKI, R. (2020). Combining machine learning and social network analysis to reveal the organizational structures. *Appl. Sci.* **10** 1699.

ONNELA, J.-P., SARAMÄKI, J., HYVÖNEN, J., SZABÓ, G., LAZER, D., KASKI, K., KERTÉSZ, J. and BARABÁSI, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **104** 7332–7336.

PAGE, L., BRIN, S., MOTWANI, R. and WINOGRAD, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab.

PALUS, S., BRODKA, P. and KAZIENKO, P. (2011). Evaluation of organization structure based on email interactions. *Int. J. Knowl. Soc. Res.* **2** 1–13.

PRIM, R. C. (1957). Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* **36** 1389–1401.

ROBBINS, S. P. (2004). *Organizational Theory*: *Structure*, *Design and Applications*.

ROWE, R., CREAMER, G., HERSHKOP, S. and STOLFO, S. J. (2007). Automated social hierarchy detection through email network analysis. In *Proceedings of the* 9*th WebKDD and* 1*st SNA-KDD* 2007 *Workshop on Web Mining and Social Network Analysis* 109–117.

SHETTY, J. and ADIBI, J. (2004). The Enron email dataset database schema and brief statistical report. Information sciences institute technical report, Univ. Southern California **4** 120–128.

SIMS, B. H., SINITSYN, N. and EIDENBENZ, S. J. (2014). Hierarchical and matrix structures in a large organizational email network: Visualization and modeling approaches. In *Social Network Analysis—Community Detection and Evolution* 27–43. Springer, Berlin.

SQUARTINI, T., PICCIOLO, F., RUZZENENTI, F. and GARLASCHELLI, D. (2013). Reciprocity of weighted networks. *Sci. Rep.* **3** 1–9.

TATTI, N. (2014). Faster way to agony. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 163–178. Springer, Berlin.

TICHY, N. M., TUSHMAN, M. L. and FOMBRUN, C. (1979). Social network analysis for organizations. *Acad. Manag. Rev.* **4** 507–519.

WANG, Y., ILIOFOTOU, M., FALOUTSOS, M. and WU, B. (2013). Analyzing communication interaction networks (CINs) in enterprises and inferring hierarchies. *Comput. Netw.* **57** 2147–2158.

ZHANG, C., HURST, W. B., LENIN, R. B., YURUK, N. and RAMASWAMY, S. (2009). Analyzing organizational structures using social network analysis. In *Advances in Enterprise Engineering III* 143–156. Springer, Berlin.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3024–3049
https://doi.org/10.1214/24-AOAS1921
This research was funded, in whole or in part, by [UK Economic and Social Research Council, ES/P000118/1].

MODELLING CORRELATION MATRICES IN MULTIVARIATE DATA, WITH APPLICATION TO RECIPROCITY AND COMPLEMENTARITY OF CHILD-PARENT EXCHANGES OF SUPPORT

BY SILIANG ZHANG[1,a], JOUNI KUHA[2,b] AND FIONA STEELE[2,c]

[1]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, School of Statistics, East China Normal University,* [a]*slzhang@fem.ecnu.edu.cn*

[2]*Department of Statistics, London School of Economics and Political Science,* [b]*j.kuha@lse.ac.uk,* [c]*f.a.steele@lse.ac.uk*

We define a model for the joint distribution of multiple continuous latent variables, which includes a model for how their correlations depend on explanatory variables. This is motivated by and applied to social scientific research questions in the analysis of intergenerational help and support within families, where the correlations describe reciprocity of help between generations and complementarity of different kinds of help. We propose an MCMC procedure for estimating the model which maintains the positive definiteness of the implied correlation matrices and describe theoretical results which justify this approach and facilitate efficient implementation of it. The model is applied to data from the UK Household Longitudinal Study to analyse exchanges of practical and financial support between adult individuals and their noncoresident parents.

REFERENCES

ALBERTINI, M., KOHLI, M. and VOGEL, C. (2007). Intergenerational transfers of time and money in European families: Common patterns—different regimes? *J. Eur. Soc. Policy* **17** 319–334.

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist.* **1** 135–141. MR0331612

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882 https://doi.org/10.1007/s11222-008-9110-y

ARCHAKOV, I. and HANSEN, P. R. (2021). A new parametrization of correlation matrices. *Econometrica* **89** 1699–1715. MR4325198 https://doi.org/10.3982/ecta16910

ATTIAS-DONFUT, C., OGG, J. and WOLFF, F.-C. (2005). European patterns of intergenerational financial and time transfers. *Eur. J. Ageing* **2** 161–173. https://doi.org/10.1007/s10433-005-0008-7

BAKK, Z. and KUHA, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika* **83** 871–892. MR3875886 https://doi.org/10.1007/s11336-017-9592-7

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BENGTSON, V. L. (2001). Beyond the nuclear family: The increasing importance of multigenerational bonds. *J. Marriage Fam.* **63** 1–16.

BONSANG, E. (2007). How do middle-aged children allocate time and money transfers to their older parents in Europe? *Empirica* **34** 171–188.

CHEN, Z. and DUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59** 762–769. MR2025100 https://doi.org/10.1111/j.0006-341X.2003.00089.x

CHENG, Y. P., BIRDITT, K. S., ZARIT, S. H. and FINGERMAN, K. L. (2015). Young adults' provision of support to middle-aged parents. *J. Gerontol., Ser. B* **70** 407–416.

CHIB, S. and GREENBERG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

CHIU, T. Y. M., LEONARD, T. and TSUI, K.-W. (1996). The matrix-logarithmic covariance model. *J. Amer. Statist. Assoc.* **91** 198–210. MR1394074 https://doi.org/10.2307/2291396

DAVEY, A. and EGGEBEEN, D. J. (1998). Patterns of intergenerational exchange and mental health. *J. Gerontol., Ser. B, Psychol. Sci. Soc. Sci.* **53B** P86–P95.

Key words and phrases. Bayesian estimation, covariance matrix modelling, item response theory models, positive definite matrices, two-step estimation.

DEINDL, C. and BRANDT, M. (2011). Financial support and practical help between older parents and their middle-aged children. *Ageing Soc*. **31** 645–62.
EVANDROU, M., FALKINGHAM, J., GOMEZ-LEON, M. and VLACHANTONI, A. (2018). Intergenerational flows of support between parents and adult children in Britain. *Ageing Soc*. **38** 321–351.
FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061
GELMAN, A., GILKS, W. R. and ROBERTS, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751 https://doi.org/10.1214/aoap/1034625254
GHOSH, R. P., MALLICK, B. and POURAHMADI, M. (2021). Bayesian estimation of correlation matrices of longitudinal data. *Bayesian Anal.* **16** 1039–1058. MR4303878 https://doi.org/10.1214/20-BA1237
GRUNDY, E. (2005). Reciprocity in relationships: Socio-economic and health influences on intergenerational exchanges between third age parents and their adult children in Great Britain. *Br. J. Sociol.* **56** 233–255. https://doi.org/10.1111/j.1468-4446.2005.00057.x
HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504 https://doi.org/10.2307/3318737
HENRETTA, J. C., VOORHIS, M. F. V. and SOLDO, B. J. (2018). Cohort differences in parental financial help to adult children. *Demography* **55** 1567–1582. https://doi.org/10.1007/s13524-018-0687-2
HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851 https://doi.org/10.1214/07-AOAS107
HOFF, P. D. and NIU, X. (2012). A covariance regression model. *Statist. Sinica* **22** 729–753. MR2954359 https://doi.org/10.5705/ss.2010.051
HOGAN, D. P., EGGEBEEN, D. J. and CLOGG, C. C. (1993). The structure of intergenerational exchanges in American families. *Am. J. Sociol.* **98** 1428–1458.
HU, J., CHEN, Y., LENG, C. and TANG, C. Y. (2021). Regression analysis of correlations for correlated data. arXiv preprint. Available at arXiv:2109.05861.
HUININK, J., BRUDERL, J., NAUCK, B., WALPER, S., CASTIGLIONI, L. and FELDHAUS, M. (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *ZfF, Z. Fam.forsch.* (*J. Fam. Res.*) **23** 77–101.
KIM, K., ZARIT, S. H., EGGEBEEN, D. J., BIRDITT, K. S. and FINGERMAN, K. L. (2011). Discrepancies in reports of support exchanges between aging parents and their middle-aged children. *J. Gerontol., Ser. B* **66** 527–537.
KUHA, J. and BAKK, Z. (2023). Two-step estimation of latent trait models. arXiv preprint. Available at arXiv:2303.16101.
KUHA, J., ZHANG, S. and STEELE, F. (2023). Latent variable models for multivariate dyadic data with zero inflation: Analysis of intergenerational exchanges of family support. *Ann. Appl. Stat.* **17** 1521–1542. MR4582723 https://doi.org/10.1214/22-aoas1680
LESTHAEGHE, R. (2014). The second demographic transition: A concise overview of its development. *Proc. Natl. Acad. Sci. USA* **111** 18112–18115.
LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456 https://doi.org/10.1093/biomet/91.1.1
LITWIN, H. (2004). Intergenerational exchange and mental health in later-life–the case of older Jewish Israelis. *Aging Ment Health* **8** 196–200. https://doi.org/10.1080/13607860410001669723
LITWIN, H., VOGEL, C., KÜNEMUND, H. and KOHLI, M. (2008). The balance of intergenerational exchange: Correlates of net transfers in Germany and Israel. *Eur. J. Ageing* **5** 91–102. https://doi.org/10.1007/s10433-008-0079-3
LUO, R. and PAN, J. (2022). Conditional generalized estimating equations of mean-variance-correlation for clustered data. *Comput. Statist. Data Anal.* **168** Paper No. 107386, 15. MR4339572 https://doi.org/10.1016/j.csda.2021.107386
MANDERMAKERS, J. J. and DYKSTRA, P. A. (2008). Discrepancies in parents' and adult child's reports of support and contact. *J. Marriage Fam.* **70** 495–506.
MUDRAZIJA, S. (2016). Public transfers and the balance of intergenerational family support in Europe. *Eur. Soc.* **18** 336–35.
MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Amer. Statist. Assoc.* **108** 656–665. MR3174649 https://doi.org/10.1080/01621459.2012.762328
MUTHÉN, L. K. and MUTHÉN, B. (2010). *Mplus User's Guide*, 6th ed. Muthén & Muthén, Los Angeles, CA.
PAN, J. and MACKENZIE, G. (2006). Regression models for covariance structures in longitudinal studies. *Stat. Model.* **6** 43–57. MR2226784 https://doi.org/10.1191/1471082X06st105oa
PAN, J. and PAN, Y. (2017). jmcm: An R package for joint mean-covariance modeling of longitudinal data. *J. Stat. Softw.* **82** 1–29.

PINHEIRO, J. C. and BATES, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Stat. Comput.* **6** 289–296.
POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786 https://doi.org/10.1093/biomet/86.3.677
POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94** 1006–1013. MR2376812 https://doi.org/10.1093/biomet/asm073
POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. MR2917961 https://doi.org/10.1214/11-STS358
ROSSEEL, Y. and LOH, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychol. Methods*. Advance Online Publication. https://dx.doi.org/10.1037/met0000503
ROUSSEEUW, P. J. and MOLENBERGHS, G. (1994). The shape of correlation matrices. *Amer. Statist.* **48** 276–279. MR1321893 https://doi.org/10.2307/2684832
SHAPIRO, A. (2004). Revisiting the generation gap: Exploring the relationships of parent/adult-child dyads. *Int. J. Aging Hum. Dev.* **58** 127–146. https://doi.org/10.2190/EVFK-7F2X-KQNV-DH58
SILVERSTEIN, M. and BENGTSON, V. L. (1997). Intergenerational solidarity and the structure of adult child-parent relationships in American families. *Am. J. Sociol.* **103** 429–460.
SILVERSTEIN, M., CONROY, S. J., WANG, H., GIARRUSSO, R. and BENGTSON, V. L. (2002). Reciprocity in parent-child relations over the adult life course. *J. Gerontol., Ser. B, Psychol. Sci. Soc. Sci.* **57B** S3–S13.
STEELE, F. and GRUNDY, E. (2021). Random effects dynamic panel models for unequally spaced multivariate categorical repeated measures: An application to child-parent exchanges of support. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 3–23. MR4204935 https://doi.org/10.1111/rssc.12446
STEELE, F., ZHANG, S., GRUNDY, E. and BURCHARDT, T. (2024). Longitudinal analysis of exchanges of support between parents and children in the UK. *J. Roy. Statist. Soc. Ser. A* **187** 279–304. MR4745593 https://doi.org/10.1093/jrsssa/qnad110
STONE, J., BERRINGTON, A. and FALKINGHAM, J. (2011). The changing determinants of UK young adults' living arrangements. *Demogr. Res.* **25** 629–666.
SUITOR, J. J., GILLIGAN, M., PILLEMER, K., FINGERMAN, K. L., KIM, K., SILVERSTEIN, M. and BENGTSON, V. L. (2017). Applying within-family differences approaches to enhance understanding of the complexity of intergenerational relations. *J. Gerontol., Ser. B* **73** 40–53.
TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1728. MR1329166 https://doi.org/10.1214/aos/1176325750
TIERNEY, L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice. Interdiscip. Statist.* 59–74. CRC Press, London. MR1397968
UNIVERSITY OF ESSEX, INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH, NATCEN SOCIAL RESEARCH AND KANTAR PUBLIC (2019). Understanding Society: Waves 1–9, 2009-2016 and Harmonised BHPS: Waves 1–18, 1991–2009. [data collection], 12th ed. ed. University of Essex, Institute for Social and Economic Research. UK, Data Service. SN: 6614.
WANG, Y. and DANIELS, M. J. (2013). Bayesian modeling of the dependence in longitudinal data via partial autocorrelations and marginal variances. *J. Multivariate Anal.* **116** 130–140. MR3049896 https://doi.org/10.1016/j.jmva.2012.11.010
WILDING, G. E., CAI, X., HUTSON, A. and YU, Z. (2011). A linear model-based test for the heterogeneity of conditional correlations. *J. Appl. Stat.* **38** 2355–2366. MR2843263 https://doi.org/10.1080/02664763.2011.559201
WONG, F., CARTER, C. K. and KOHN, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90** 809–830. MR2024759 https://doi.org/10.1093/biomet/90.4.809
YAN, J. and FINE, J. (2007). Estimating equations for association structures. *Stat. Med.* **23** 859–874.
ZHANG, S., KUHA, J. and STEELE, F. (2024). Supplement to "Modelling correlation matrices in multivariate data, with application to reciprocity and complementarity of child-parent exchanges of support." https://doi.org/10.1214/24-AOAS1921SUPPA, https://doi.org/10.1214/24-AOAS1921SUPPB
ZHANG, W. and LENG, C. (2012). A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika* **99** 141–150. MR2899669 https://doi.org/10.1093/biomet/asr068
ZHANG, W., LENG, C. and TANG, C. Y. (2015). A joint modelling approach for longitudinal studies. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 219–238. MR3299406 https://doi.org/10.1111/rssb.12065
ZOU, T., LAN, W., LI, R. and TSAI, C.-L. (2022). Inference on covariance-mean regression. *J. Econometrics* **230** 318–338. MR4466727 https://doi.org/10.1016/j.jeconom.2021.05.004
ZOU, T., LAN, W., WANG, H. and TSAI, C.-L. (2017). Covariance regression analysis. *J. Amer. Statist. Assoc.* **112** 266–281. MR3646570 https://doi.org/10.1080/01621459.2015.1131699
ZWIERNIK, P., UHLER, C. and RICHARDS, D. (2017). Maximum likelihood estimation for linear Gaussian covariance models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1269–1292. MR3689318 https://doi.org/10.1111/rssb.12217

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3050–3070
https://doi.org/10.1214/24-AOAS1922

BAYESIAN HIDDEN MARKOV MODEL FOR NATURAL HISTORY OF COLORECTAL CANCER: HANDLING MISCLASSIFIED OBSERVATIONS, VARYING OBSERVATION SCHEMES AND UNOBSERVED DATA

BY AAPELI NEVALA[1,a], SIRPA HEINÄVAARA[1,b], TYTTI SARKEALA[1,c] AND SANGITA KULATHINAL[2,d]

[1]*Mass Screening Registry, Finnish Cancer Registry,* [a]*aapeli.nevala@cancer.fi,* [b]*sirpa.heinavaara@cancer.fi,* [c]*tytti.sarkeala@cancer.fi*

[2]*Department of Mathematics and Statistics, University of Helsinki,* [d]*sangita.kulathinal@helsinki.fi*

Statistical modelling of individual-level event history data arising from varying observation schemes is a challenging problem, particularly due to unobserved and possibly misclassified individual states. Commonly used approaches rely on the hidden Markov models (HMM) to incorporate true underlying states. Each approach needs to account for the underlying data generating process and related external information and requires assumptions for estimation. This article develops a Bayesian HMM for natural history of colorectal cancer (CRC), combining data on latent disease states from randomised screening study and on observed clinical cancers from the population-based cancer registry. With our modelling approach and study design, we are able to provide estimates for latent state occupancy probabilities not only for screening-attenders but also for the control group and those who never attended screening—despite data on latent states only existing for the attenders. We use simulation-based calibration to ensure that posterior distributions can be reliably estimated despite the challenges brought in by the sampling scheme. We apply Bayesian computation to obtain posterior estimates of the quantities of interest. Two algorithms, Hamiltonian Monte Carlo (HMC) and Automatic Differentiation Variational Inference (ADVI), are applied and compared, first by using simulated data and then with a real data set. The modelling workflow can be applied for different cancer screening programmes and datasets which typically have similar challenges.

REFERENCES

AALEN, O. O., BORGAN, Ø. and GJESSING, H. K. (2008). *Survival and Event History Analysis*: *A Process Point of View*. *Statistics for Biology and Health*. Springer, New York. MR2449233 https://doi.org/10.1007/978-0-387-68560-1

ANNUAL STATISTICS (2021). Finnish Cancer Registry Annual Statistics. Available at https://tilastot.syoparekisteri.fi/syovat (data from 2021-10-04, version 2021-10-22-002).

BETANCOURT, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. Available at arXiv:1701.02434 [stat].

BLOM, J., YIN, L., LIDÉN, A., DOLK, A., JEPPSSON, B., PÅHLMAN, L., HOLMBERG, L. and NYRÉN, O. (2008). A 9-year follow-up study of participants and nonparticipants in sigmoidoscopy screening: Importance of self-selection. *Cancer Epidemiol. Biomark. Prev.* **17** 1163–1168.

BRENNER, H., ALTENHOFEN, L., KATALINIC, A., LANSDORP-VOGELAAR, I. and HOFFMEISTER, M. (2011). Sojourn time of preclinical colorectal cancer by sex and age: Estimates from the German national screening colonoscopy database. *Amer. J. Epidemiol.* **174** 1140–1146. https://doi.org/10.1093/aje/kwr188

BRENNER, H., HAUG, U. and HUNDT, S. (2010). Sex differences in performance of fecal occult blood testing. *Amer. J. Gastroenterol.* **105** 2457–2464. https://doi.org/10.1038/ajg.2010.301

BRENNER, H., HOFFMEISTER, M., STEGMAIER, C., BRENNER, G., ALTENHOFEN, L. and HAUG, U. (2007). Risk of progression of advanced adenomas to colorectal cancer by age and sex: Estimates based on 840,149 screening colonoscopies. *Gut* **56** 1585–1589.

Key words and phrases. Bayesian statistics, multistate models, event-history analysis, hidden Markov model, Stan, cancer screening, cancer epidemiology.

BÜRKNER, P.-C., GABRY, J., KAY, M. and VEHTARI, A. (2022). posterior: Tools for Working with Posterior Distributions. R package version 1.3.1.
CARPENTER, B. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.*
COOK, R. J. and LAWLESS, J. F. (2018). *Multistate Models for the Analysis of Life History Data*, 2nd ed. CRC Press, Boca Raton. https://doi.org/10.1201/9781315119731
CORLEY, D. A., JENSEN, C. D., MARKS, A. R., ZHAO, W. K., DE BOER, J., LEVIN, T. R., DOUBENI, C., FIREMAN, B. H. and QUESENBERRY, C. P. (2013). Variation of adenoma prevalence by age, sex, race, and colon location in a large population: Implications for screening and quality programs. *Clin. Gastroenterol. Hepatol.* **11** 172–180.
FEARON, E. F. and BERT, V. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61** 759–767.
FERLITSCH, M., REINHART, K., PRAMHAS, S., WIENER, C., GAL, O., BANNERT, C., HASSLER, M., KOZBIAL, K., DUNKLER, D. et al. (2011). Sex-specific prevalence of adenomas, advanced adenomas, and colorectal cancer in individuals undergoing screening colonoscopy. *JAMA* **306** 1352–1358.
GABRY, J. and ČEŠNOVAR, R. (2021). Cmdstanr: R Interface to 'CmdStan'. Available at https://mc-stan.org/cmdstanr.
GELMAN, A., CARLIN, J., STERN, H., DUNSON, D., VEHTARI, A. and DB, R. (2013). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton. https://doi.org/10.1201/b16018
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. https://doi.org/10.48550/ARXIV.2011.01808
GRINSZTAJN, L., SEMENOVA, E., MARGOSSIAN, C. C. and RIOU, J. (2021). Bayesian workflow for disease transmission modeling in Stan. *Stat. Med.* **40** 6209–6234. MR4339396 https://doi.org/10.1002/sim.9164
HAKAMA, M., POKHREL, A., MALILA, N. and HAKULINEN, T. (2015). Sensitivity, effect and overdiagnosis in screening for cancers with detectable pre-invasive phase. *Int. J. Cancer* **136** 928–935. https://doi.org/10.1002/ijc.29053
HARDCASTLE, J. D., CHAMBERLAIN, J. O., ROBINSON, M. H., MOSS, S. M., AMAR, S. S., BALFOUR, T. W., JAMES, P. D. and MANGHAM, C. M. (1996). Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. *Lancet* **348** 1472–1477. https://doi.org/10.1016/S0140-6736(96)03386-7
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
JØRGENSEN, O. D., KRONBORG, O. and FENGER, C. (2002). A randomised study of screening for colorectal cancer using faecal occult blood testing: Results after 13 years and seven biennial screening rounds. *Gut* **50** 29–32.
KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867
KLAUSCH, T., AKWIWU, E. U., VAN DE WIEL, M. A., COUPÉ, V. M. H. and BERKHOF, J. (2023). A Bayesian accelerated failure time model for interval censored three-state screening outcomes. *Ann. Appl. Stat.* **17** 1285–1306. MR4582713 https://doi.org/10.1214/22-aoas1669
KOSKENVUO, L., MALILA, N., NIEMI, J., MIETTINEN, J., HEIKKINEN, S. and SALLINEN, V. (2019). Sex differences in faecal occult blood test screening for colorectal cancer. *Br. J. Surg.* **106** 436–447.
KUCUKELBIR, A., TRAN, D., RANGANATH, R., GELMAN, A. and BLEI, D. M. (2017). Automatic differentiation variational inference. *J. Mach. Learn. Res.* **18** Paper No. 14, 45. MR3634881
KUNTZ, K. M., LANSDORP-VOGELAAR, I., RUTTER, C. M., KNUDSEN, A. B., VAN BALLEGOOIJEN, M., SAVARINO, J. E., FEUER, E. J. and ZAUBER, A. G. (2011). A systematic comparison of microsimulation models of colorectal cancer: The role of assumptions about adenoma progression. *Med. Decis. Mak.* **31** 530–539. https://doi.org/10.1177/0272989X11408730
LANGE, J. M., HUBBARD, R. A., INOUE, L. Y. T. and MININ, V. N. (2015). A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* **71** 90–101. MR3335353 https://doi.org/10.1111/biom.12252
LEINONEN, M. K., MIETTINEN, J., HEIKKINEN, S., PITKÄNIEMI, J. and MALILA, N. (2017). Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. *Eur. J. Cancer* **77** 31–39. https://doi.org/10.1016/j.ejca.2017.02.017
LUO, Y., STEPHENS, D. A., VERMA, A. and BUCKERIDGE, D. L. (2021). Bayesian latent multi-state modeling for nonequidistant longitudinal electronic health records. *Biometrics* **77** 78–90. MR4229722 https://doi.org/10.1111/biom.13261
MALILA, N., ANTTILA, A. and HAKAMA, M. (2005). Colorectal cancer screening in Finland: Details of the national screening programme implemented in Autumn 2004. *J. Med. Screen.* **12** 28–32. https://doi.org/10.1258/0969141053279095
MANDEL, J. S., BOND, J. H., CHURCH, T. R., SNOVER, D. C., BRADLEY, G. M., SCHUMAN, L. M. and EDERER, F. (1993). Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota colon cancer control study. *N. Engl. J. Med.* **328** 1365–1371.

NEVALA, A., HEINÄVAARA, S., SARKEALA, T. and KULATHINAL, S. (2024). Supplement to "Bayesian hidden Markov model for natural history of colorectal cancer: handling misclassified observations, varying observation schemes and unobserved data." https://doi.org/10.1214/24-AOAS1922SUPPA, https://doi.org/10.1214/24-AOAS1922SUPPB

PANKAKOSKI, M., HEINÄVAARA, S., ANTTILA, A. and SARKEALA, T. (2020). Differences in cervical test coverage by age, socioeconomic status, ethnic origin and municipality type—a nationwide register-based study. *Prev. Med.* **139** 106219. https://doi.org/10.1016/j.ypmed.2020.106219

PHILLIPS, C. J. and SCHOEN, R. E. (2020). Screening for colorectal cancer in the age of simulation models: A historical lens. *Gastroenterology* **159** 1201–1204. https://doi.org/10.1053/j.gastro.2020.07.010

PITKÄNIEMI, J., MALILA, N., TANSKANEN, T., DEGERLUND, H., HEIKKINEN, S. and SEPPÄ, K. (2021). Cancer in Finland 2019.

RUTTER, C. M., OZIK, J., DEYOREO, M. and COLLIER, N. (2019). Microsimulation model calibration using incremental mixture approximate Bayesian computation. *Ann. Appl. Stat.* **13** 2189–2212. MR4037427 https://doi.org/10.1214/19-aoas1279

SILVA-ILLANES, N. and ESPINOZA, M. (2018). Critical analysis of Markov models used for the economic evaluation of colorectal cancer screening: A systematic review. *Value Health* **21** 858–873.

SISSON, S. A., FAN, Y. and BEAUMONT, M. A. (2019). Overview of ABC. In *Handbook of Approximate Bayesian Computation. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 3–54. CRC Press, Boca Raton, FL. MR3889278

STAN DEVELOPMENT TEAM (2023). Stan Modeling Language Users Guide and Reference Manual, 2.28.

SUNG, H., FERLAY, J., SIEGEL, R. L., LAVERSANNE, M., SOERJOMATARAM, I., JEMAL, A. and BRAY, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71** 209–249. https://doi.org/10.3322/caac.21660

SUNG, N. Y., JUN, J. K., KIM, Y. N., JUNG, I., PARK, S., KIM, G. R. and NAM, C. M. (2019). Estimating age group-dependent sensitivity and mean sojourn time in colorectal cancer screening. *J. Med. Screen.* **26** 19–25.

TALTS, S., BETANCOURT, M., SIMPSON, D., VEHTARI, A. and GELMAN, A. (2020). Validating Bayesian Inference Algorithms with Simulation-Based Calibration.

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

WILLIAMS, J. P., STORLIE, C. B., THERNEAU, T. M., JACK, C. R. JR. and HANNIG, J. (2020). A Bayesian approach to multistate hidden Markov models: Application to dementia progression. *J. Amer. Statist. Assoc.* **115** 16–31. MR4078442 https://doi.org/10.1080/01621459.2019.1594831

WINAWER, S. J. and ZAUBER, A. G. (2002). The advanced adenoma as the primary target of screening. *Gastrointest. Endosc. Clin. N. Amer.* **12** 1–v. https://doi.org/10.1016/s1052-5157(03)00053-9

WINAWER, S. J., ZAUBER, A. G., HO, M. N., O'BRIEN, M. J., GOTTLIEB, L. S., STERNBERG, S. S., WAYE, J. D., SCHAPIRO, M., BOND, J. H. et al. (1993). Prevention of colorectal cancer by colonoscopic polypectomy. The national polyp study workgroup. *N. Engl. J. Med.* **329** 1977–1981. https://doi.org/10.1056/NEJM199312303292701

YAMANE, L., SCAPULATEMPO-NETO, C., REIS, R. M. and GUIMARÃES, D. P. (2014). Serrated pathway in colorectal carcinogenesis. *World J. Gastroenterol.* **20** 2634–2640. https://doi.org/10.3748/wjg.v20.i10.2634

YAO, Y., VEHTARI, A., SIMPSON, D. and GELMAN, A. (2018). Yes, but did it work?: Evaluating variational inference. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 5581–5590. PMLR, US.

ZHENG, W. and RUTTER, C. M. (2012). Estimated mean sojourn time associated with hemoccult SENSA for detection of proximal and distal colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **21** 1722–1730.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3071–3090
https://doi.org/10.1214/24-AOAS1924

ASSESSING MARINE MAMMAL ABUNDANCE: A NOVEL DATA FUSION

BY ERIN M. SCHLIEP[1,a], ALAN E. GELFAND[2,b], CHRISTOPHER W. CLARK[3,c], CHARLES A. MAYO[4,d], BRIGID MCKENNA[4,e], SUSAN E. PARKS[5,f], TINA M. YACK[6,g] AND ROBERT S. SCHICK[7,h]

[1]*Department of Statistics, North Carolina State University,* [a]*emschliep@ncsu.edu*
[2]*Department of Statistical Science, Duke University,* [b]*alan@duke.edu*
[3]*K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University,* [c]*cwc2@cornell.edu*
[4]*Right Whale Ecology Program, Center for Coastal Studies,* [d]*c.mayoiii@comcast.net,* [e]*bmckenna@coastalstudies.org*
[5]*Biology Department, Syracuse University,* [f]*sparks@syr.edu*
[6]*Nicholas School of the Environment, Duke University,* [g]*tina.yack@duke.edu*
[7]*Southall Environmental Associates, Inc.* [h]*robschick@sea-inc.net*

Marine mammals are increasingly vulnerable to human disturbance and climate change. Their diving behavior leads to limited visual access during data collection, making studying the abundance and distribution of marine mammals challenging. In theory, using data from more than one observation modality should lead to better informed predictions of abundance and distribution. With focus on North Atlantic right whales, we consider the fusion of two data sources to inform about their abundance and distribution. The first source is aerial distance sampling, which provides the spatial locations of whales detected in the region. The second source is passive acoustic monitoring (PAM), returning calls received at hydrophones placed on the ocean floor. Due to limited time on the surface and detection limitations arising from sampling effort, aerial distance sampling only provides a partial realization of locations. With PAM we never observe numbers or locations of individuals. To address these challenges, we develop a novel *thinned* point pattern data fusion. Our approach leads to improved inference regarding abundance and distribution of North Atlantic right whales throughout Cape Cod Bay, Massachusetts in the U.S. We demonstrate performance gains of our approach compared to that from a single source through both simulation and real data.

REFERENCES

BUCKLAND, S. T. (2006). Point-transect surveys for songbirds: Robust methodologies. *Auk* **123** 345–357. https://doi.org/10.1642/0004-8038(2006)123[345:PSFSRM]2.0.CO;2

BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P., LAAKE, J. L., BORCHERS, D. L. and THOMAS, L. (2001). *Introduction to Distance Sampling*: *Estimating Abundance of Biological Populations*. Oxford Univ. Press, London. MR1263023

BUCKLAND, S. T., ANDERSON, D. R., BURNHAM, K. P. and THOMAS, L. (2007). *Advanced Distance Sampling*. Oxford Univ. Press, London.

BUCKLAND, S. T., BORCHERS, D. L., MARQUES, T. A. and FEWSTER, R. M. (2023). Wildlife population assessment: Changing priorities driven by technological advances. *J. Stat. Theory Pract.* **17** Paper No. 20, 22. MR4546848 https://doi.org/10.1007/s42519-023-00319-6

BUCKLAND, S. T., REXSTAD, E. A., MARQUES, T. A. and OEDEKOVEN, C. S. (2015). *Distance Sampling*: *Methods and Applications*: *Methods in Statistical Ecology*. Springer, Berlin. https://doi.org/10.1007/978-3-319-19219-2

CHAKRABORTY, A., GELFAND, A. E., WILSON, A. M., LATIMER, A. M. and SILANDER, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 757–776. MR2844854 https://doi.org/10.1111/j.1467-9876.2011.01023.x

CLARK, C. W. (1982). The acoustic repertoire of the southern right whale, a quantitative analysis. *Anim. Behav.* **30** 1060–1071. https://doi.org/10.1016/s0003-3472(82)80196-6

Key words and phrases. Data assimilation, data integration, Bayesian hierarchical modeling, North Atlantic right whales, point pattern data, thinning.

CLARK, C. W., BROWN, M. W. and CORKERON, P. (2010). Visual and acoustic surveys for North Atlantic right whales, Eubalaena glacialis, in Cape Cod Bay, Massachusetts, 2001–2005: Management implications. *Mar. Mamm. Sci.* **26** 837–854. https://doi.org/10.1111/j.1748-7692.2010.00376.x

COUTINHO, R. W. and BOUKERCHE, A. (2021). North Atlantic right whales preservation: A new challenge for Internet of underwater things and smart ocean-based systems. *IEEE Instrum. Meas. Mag.* **24** 61–67.

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.

DORAZIO, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr.* **23** 1472–1484.

DOVERS, E., POPOVIC, G. C. and WARTON, D. I. (2024). A fast method for fitting integrated species distribution models. *Methods Ecol. Evol.* **15** 191–203.

FARR, M. T., GREEN, D. S., HOLEKAMP, K. E. and ZIPKIN, E. F. (2021). Integrating distance sampling and presence-only data to estimate species abundance. *Ecology* **102** e03204. https://doi.org/10.1002/ecy.3204

FINLEY, A. O., BANERJEE, S. and BASSO, B. (2011). Improving crop model inference through Bayesian melding with spatially varying parameters. *J. Agric. Biol. Environ. Stat.* **16** 453–474. MR2862293 https://doi.org/10.1007/s13253-011-0070-x

FITHIAN, W., ELITH, J., HASTIE, T. and KEITH, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **6** 424–438. https://doi.org/10.1111/2041-210X.12242

FOLEY, K. M. and FUENTES, M. (2008). A statistical framework to combine multivariate spatial data and physical models for hurricane surface wind prediction. *J. Agric. Biol. Environ. Stat.* **13** 37–59. MR2423075 https://doi.org/10.1198/108571108X276473

FRANKLIN, K. J., COLE, T. V. N., CHOLEWIAK, D. M., DULEY, P. A., CROWE, L. M., HAMILTON, P. K., KNOWLTON, A. R., TAGGART, C. T. and JOHNSON, H. D. (2022). Using sonobuoys and visual surveys to characterize North Atlantic right whale (Eubalaena glacialis) calling behavior in the Gulf of St. Lawrence. *Endanger. Species Res.* **49** 159–174. https://doi.org/10.3354/esr01208

FRASIER, K. E., GARRISON, L. P., SOLDEVILLA, M. S., WIGGINS, S. M. and HILDEBRAND, J. A. (2021). Cetacean distribution models based on visual and passive acoustic data. *Sci. Rep.* **11** 8240. https://doi.org/10.1038/s41598-021-87577-1

FRISTRUP, K. M. and CLARK, C. W. (1997). Combining visual and acoustic survey data to enhance density estimation. *Rep. Int. Whal. Comm.* **47** 933–936.

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

GANLEY, L. C., BRAULT, S. and MAYO, C. A. (2019). What we see is not what there is: Estimating North Atlantic right whale Eubalaena glacialis local abundance. *Endanger. Species Res.* **38** 101–113. https://doi.org/10.3354/esr00938

GELFAND, A. E. and SCHLIEP, E. M. (2018). *Bayesian Inference and Computing for Spatial Point Patterns. NSF-CBMS Regional Conference Series in Probability and Statistics* **10**. IMS, Beachwood, OH, Amer. Statist. Assoc., Alexandria, VA. MR3890052

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

HATCH, L. T., CLARK, C. W., VAN PARIJS, S. M., FRANKEL, A. S. and PONIRAKIS, D. W. (2012). Quantifying loss of acoustic communication space for right whales in and around a U.S. National Marine Sanctuary. *Conserv. Biol.* **26** 983–994. https://doi.org/10.1111/j.1523-1739.2012.01908.x

HEDLEY, S. L. and BUCKLAND, S. T. (2004). Spatial models for line transect sampling. *J. Agric. Biol. Environ. Stat.* **9** 181–199. https://doi.org/10.1198/1085711043578

HUDAK, C. A., STAMIESZKIN, K. and MAYO, C. A. (2023). North Atlantic right whale Eubalaena glacialis prey selection in Cape Cod Bay. *Endanger. Species Res.* **51** 15–29. https://doi.org/10.3354/esr01240

ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns. Statistics in Practice*. Wiley, Chichester. MR2384630

ISAAC, N. J., JARZYNA, M. A., KEIL, P., DAMBLY, L. I., BOERSCH-SUPAN, P. H., BROWNING, E., FREEMAN, S. N., GOLDING, N., GUILLERA-ARROITA, G. et al. (2020). Data integration for large-scale models of species distributions. *Trends Ecol. Evol.* **35** 56–67.

JOHNSON, D. S., LAAKE, J. L. and VER HOEF, J. M. (2010). A model-based approach for making ecological inference from distance sampling data. *Biometrics* **66** 310–318. MR2756719 https://doi.org/10.1111/j.1541-0420.2009.01265.x

LIU, Z., LE, N. D. and ZIDEK, J. V. (2011). An empirical assessment of Bayesian melding for mapping ozone pollution. *Environmetrics* **22** 340–353. MR2843389 https://doi.org/10.1002/env.1054

MARQUES, T. A., MUNGER, L., THOMAS, L., WIGGINS, S. and HILDEBRAND, J. A. (2011). Estimating North Pacific right whale Eubalaena japonica density using passive acoustic cue counting. *Endanger. Species Res.* **13** 163–172. https://doi.org/10.3354/esr00325

MARQUES, T. A., THOMAS, L., MARTIN, S. W., MELLINGER, D. K., WARD, J. A., MORETTI, D. J., HARRIS, D. and TYACK, P. L. (2013). Estimating animal population density using passive acoustics. *Biol. Rev.* **88** 287–309. https://doi.org/10.1111/brv.12001

MARTINO, S., PACE, D. S., MORO, S., CASOLI, E., VENTURA, D., FRACHEA, A., SILVESTRI, M., ARCANGELI, A., GIACOMINI, G. et al. (2021). Integration of presence-only data from several sources: A case study on dolphins' spatial distribution. *Ecography* **44** 1533–1543.

MAYO, C. A., GANLEY, L., HUDAK, C. A., BRAULT, S., MARX, M. K., BURKE, E. and BROWN, M. W. (2018). Distribution, demography, and behavior of North Atlantic right whales (Eubalaena Glacialis) in Cape Cod Bay, Massachusetts, 1998-2013. *Mar. Mamm. Sci.* **34** 979–996. https://doi.org/10.1111/mms.12511

MILLER, D. L., BURT, M. L., REXSTAD, E. A. and THOMAS, L. (2013). Spatial models for distance sampling data: Recent developments and future directions. *Methods Ecol. Evol.* **4** 1001–1010. https://doi.org/10.1111/2041-210X.12105

NGUYEN, H., CRESSIE, N. and BRAVERMAN, A. (2012). Spatial statistical data fusion for remote sensing applications. *J. Amer. Statist. Assoc.* **107** 1004–1018. MR3010886 https://doi.org/10.1080/01621459.2012.694717

NOWACEK, D. P., CHRISTIANSEN, F., BEJDER, L., GOLDBOGEN, J. A. and FRIEDLAENDER, A. S. (2016). Studying cetacean behaviour: New technological approaches and conservation applications. *Anim. Behav.* **120** 235–244. https://doi.org/10.1016/j.anbehav.2016.07.019

PALMER, K. J., TABBUTT, S., GILLESPIE, D., TURNER, J., KING, P., TOLLIT, D., THOMPSON, J. and WOOD, J. (2022a). Evaluation of a coastal acoustic buoy for cetacean detections, bearing accuracy and exclusion zone monitoring. *Methods Ecol. Evol.* **13** 2491–2502. https://doi.org/10.1111/2041-210X.13973

PALMER, K. J., WU, G.-M., CLARK, C. and KLINCK, H. (2022b). Accounting for the Lombard effect in estimating the probability of detection in passive acoustic aurveys: Applications for single sensor mitigation and monitoring. *J. Acoust. Soc. Amer.* **151** 67–79. https://doi.org/10.1121/10.0009168

PARKS, S. E., SEARBY, A., CÉLÉRIER, A., JOHNSON, M. P., NOWACEK, D. P. and TYACK, P. L. (2011). Sound production behavior of individual North Atlantic right whales: Implications for passive acoustic monitoring. *Endanger. Species Res.* **15** 63–76.

RAFTERY, A. E., ZEH, J. E., YANG, Q. and STYER, P. E. (1990). Bayes empirical Bayes interval estimation of bowhead whale, Balaena mysticetus, population size based upon the 1986 combined visual and acoustic census off Point Barrow, Alaska. *Rep. Int. Whal. Comm.* **40** 393–409.

RUNDEL, C. W., SCHLIEP, E. M., GELFAND, A. E. and HOLLAND, D. M. (2015). A data fusion approach for spatial analysis of speciated $PM_{2.5}$ across time. *Environmetrics* **26** 515–525. MR3431926 https://doi.org/10.1002/env.2369

SCHLIEP, E. M., COLLINS, S. M., ROJAS-SALAZAR, S., LOTTIG, N. R. and STANLEY, E. H. (2020). Data fusion model for speciated nitrogen to identify environmental drivers and improve estimation of nitrogen in lakes. *Ann. Appl. Stat.* **14** 1651–1675. MR4194242 https://doi.org/10.1214/20-AOAS1371

SCHLIEP, E. M., GELFAND, A. E., CLARK, C. W., MAYO, C. A., MCKENNA, B., PARKS, S. E., YACK, T. M. and SCHICK, R. S. (2024). Supplement to "Assessing marine mammal abundance: a novel data fusion." https://doi.org/10.1214/24-AOAS1924SUPP

SIGOURNEY, D. B., DEANGELIS, A., CHOLEWIAK, D. and PALKA, D. (2023). Combining passive acoustic data from a towed hydrophone array with visual line transect data to estimate abundance and availability bias of sperm whales (*Physeter macrocephalus*). *PeerJ* **11** e15850. https://doi.org/10.7717/peerj.15850

SIMMONDS, E. G., JARVIS, S. G., HENRYS, P. A., ISAAC, N. J. and O'HARA, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography* **43** 1413–1422.

WATKINS, W. A. and SCHEVILL, W. E. (1972). Sound source location by arrival-times on a non-rigid three-dimensional hydrophone array. *Deep-Sea Res. Oceanogr. Abstr.* **19** 691–706. https://doi.org/10.1016/0011-7471(72)90061-7

WIKLE, C. K. and BERLINER, L. M. (2005). Combining information across spatial scales. *Technometrics* **47** 80–91. MR2099410 https://doi.org/10.1198/004017004000000572

WIKLE, C. K., MILLIFF, R. F., NYCHKA, D. and BERLINER, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc.* **96** 382–397. MR1939342 https://doi.org/10.1198/016214501753168109

YUAN, Y., BACHL, F. E., LINDGREN, F., BORCHERS, D. L., ILLIAN, J. B., BUCKLAND, S. T., RUE, H. and GERRODETTE, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. Appl. Stat.* **11** 2270–2297. MR3743297 https://doi.org/10.1214/17-AOAS1078

ZIDEK, J. V., LE, N. D. and LIU, Z. (2012). Combining data and simulated data for space-time fields: Application to ozone. *Environ. Ecol. Stat.* **19** 37–56. MR2909084 https://doi.org/10.1007/s10651-011-0172-1

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3091–3108
https://doi.org/10.1214/24-AOAS1925

BAYESIAN MODELING OF INSURANCE CLAIMS FOR HAIL DAMAGE

BY OPHÉLIA MIRALLES[1,a] AND ANTHONY C. DAVISON[2,b]

[1]*Center for Climate Systems Modeling (C2SM), ETH Zürich,* [a]*ophelia.miralles@usys.ethz.ch*
[2]*Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL),* [b]*anthony.davison@epfl.ch*

Despite its importance for insurance, there is almost no literature on statistical hail damage modeling. Statistical models for hailstorms exist, though they are generally not open-source, but no study appears to have developed a stochastic hail impact function. In this paper we use hail-related insurance claim data to build a Gaussian line process with extreme marks in order to model both the geographical footprint of a hailstorm and the damage to buildings that hailstones can cause. We build a model for the claim counts and claim values, and compare it to the use of a benchmark deterministic hail impact function. Our model proves to be better than the benchmark at capturing hail spatial patterns and allows for localized and extreme damage, which is seen in the insurance data. The evaluation of both the claim counts and value predictions shows that performance is improved compared to the benchmark, especially for extreme damage. Our model appears to be the first to provide realistic estimates for hail damage to individual buildings.

REFERENCES

AZNAR-SIGUAN, G. and BRESCH, D. N. (2019). CLIMADA v1: A global weather and climate risk assessment platform. *Geosci. Model Dev.* **12** 3085–3097. https://doi.org/10.5194/gmd-12-3085-2019

BETANCOURT, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434. https://doi.org/10.48550/arXiv.1701.02434

BORUFF, B. J., EASOZ, J. A., JONES, S. D., LANDRY, H. R., MITCHEM, J. D. and CUTTER, S. L. (2003). Tornado hazards in the United States. *Clim. Res.* **24** 103–117.

BOTZEN, W. J. W., BOUWER, L. M. and VAN DEN BERGH, J. C. J. M. (2010). Climate change and hailstorm damage: Empirical evidence and implications for agriculture and insurance. *Resour. Energy Econ.* **32** 341–362. https://doi.org/10.1016/j.reseneeco.2009.10.004

BROWN, T. M., POGORZELSKI, W. H. and GIAMMANCO, I. M. (2015). Evaluating hail damage using property insurance claims data. *Weather, Climate, and Society* **7** 197–210.

CASTRO-CAMILO, D., HUSER, R. and RUE, H. (2019). A spliced gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting. *J. Agric. Biol. Environ. Stat.* **24** 517–534. MR3996457 https://doi.org/10.1007/s13253-019-00369-z

CHANGNON, S. A. (2008). Temporal and spatial distributions of damaging hail in the continental United States. *Prog. Phys. Geogr.* **29** 341–350. https://doi.org/10.2747/0272-3646.29.4.341

CHANGNON, S. A. (2009). Tornado losses in the United States. *Natural Hazards Review* **10** 145–150. https://doi.org/10.1061/(ASCE)1527-6988(2009)10:4(145)

DANIEL, W. W. (1990). *Applied Nonparametric Statistics*, 2nd ed. Duxbury, Pacific Grove, CA.

DEEPEN, J. (2006). Schadenmodellierung extremer Hagelereignisse in Deutschland Master's thesis Westfälischen Wilhelms-Universität Münster.

GELFAND, A. E. (2000). Gibbs sampling. *J. Amer. Statist. Assoc.* **95** 1300–1304. MR1825281 https://doi.org/10.2307/2669775

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

HAUG, O., DIMAKOS, X. K., VÅRDAL, J. F., ALDRIN, M. and MEZE-HAUSKEN, E. (2011). Future building water loss projections posed by climate change. *Scand. Actuar. J.* **1** 1–20. MR2773749 https://doi.org/10.1080/03461230903266533

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.

Key words and phrases. Climate extremes, extreme value analysis, hail damage, line process, Bayesian modeling.

HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B. and HOCHREITER, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems* (*NIPS*), 4–9 *December* 2017, *Long Beach*, *CA*, *United States* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) **30** 6626–6637. Curran Associates, Red Hook, NY, United States.

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

HOHL, R., SCHIESSER, H.-H. and ALLER, D. (2002). Hailfall: The relationship between radar-derived hail kinetic energy and hail damage to buildings. *Atmos. Res.* **63** 177–207. https://doi.org/10.1016/S0169-8095(02)00059-5

JEONG, J., JUN, M. and GENTON, M. G. (2017). Spherical process models for global spatial statistics. *Statist. Sci.* **32** 501–513. MR3730519 https://doi.org/10.1214/17-STS620

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

KOPP, J., SCHRÖER, K., SCHWIERZ, C., HERING, A., GERMANN, U. and MARTIUS, O. (2023). The summer 2021 Switzerland hailstorms: Weather situation, major impacts and unique observational data. *Weather* **78** 184–191. https://doi.org/10.1002/wea.4306

LAUDAGÉ, C., DESMETTRE, S. and WENZEL, J. (2019). Severity modeling of extreme insurance claims for tariffication. *Insurance Math. Econom.* **88** 77–92. MR3973942 https://doi.org/10.1016/j.insmatheco.2019.06.002

LIU, K., LI, P. and WANG, Z. (2021). Statistical modeling of random hail impact. *Extreme Mechanics Letters* **48** 101374. https://doi.org/10.1016/j.eml.2021.101374

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (2004). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092. https://doi.org/10.1063/1.1699114

MIRALLES, O. and DAVISON, A. C (2024). Supplement to "Bayesian modeling of insurance claims for hail damage." https://doi.org/10.1214/24-AOAS1925SUPPA, https://doi.org/10.1214/24-AOAS1925SUPPB

MIRALLES, O., STEINFELD, D., MARTIUS, O. and DAVISON, A. C. (2022). Downscaling of historical wind fields over Switzerland using generative adversarial networks. *Artificial Intelligence for the Earth Systems* **1** e220018. https://doi.org/10.1175/AIES-D-22-0018.1

MÜLLER, A. (2021). 2 Milliarden Franken wegen Sturm und Hagel: 2021 wird für die Schweiz eines der teuersten Schadenjahre aller Zeiten. https://www.nzz.ch/wirtschaft/unwetter-schweiz-2021-hagel-und-sturm-kosteten-2-mrd-franken-ld.1652483?reduced=true. Accessed: 2023-06-25.

NISI, L., HERING, A., GERMANN, U. and MARTIUS, O. (2018). A 15-year hail streak climatology for the Alpine region. *Q. J. R. Meteorol. Soc.* **144** 1429–1449. https://doi.org/10.1002/qj.3286

NISI, L., MARTIUS, O., HERING, A., KUNZ, M. and GERMANN, U. (2016). Spatial and temporal distribution of hailstorms in the Alpine region: A long-term, high resolution, radar-based analysis. *Q. J. R. Meteorol. Soc.* **142** 1590–1604. https://doi.org/10.1002/qj.2771

OTTO, M. (2009). Modellierung von Hagelschäden in der Pkw-Kaskoversicherung in Deutschland Master's thesis Technische Universität Dresden.

PALDYNSKI, H. (2015). Modelling large claims in property and home insurance—extreme value analysis Master's thesis Lund Univ.

PERERA, S., LAM, N., PATHIRANA, M., ZHANG, L., RUAN, D. and GAD, E. (2018). Probabilistic modelling of forces of hail. *Nat. Hazards* **91** 133–153. https://doi.org/10.1007/s11069-017-3117-7

PORTMANN, R., SCHMID, T., VILLIGER, L., BRESCH, D. N. and CALANCA, P. (2023). Modelling crop hail damage footprints with single-polarization radar: The roles of spatial resolution, hail intensity, and cropland density. *EGUsphere* **2023** 1–29. https://doi.org/10.5194/egusphere-2023-2598

PÚČIK, T., GROENEMEIJER, P., RÄDLER, A. T., TIJSSEN, L., NIKULIN, G., PREIN, A. F., VAN MEIJGAARD, E., FEALY, R., JACOB, D. et al. (2017). Future changes in European severe convection environments in a regional climate model ensemble. *J. Climate* **30** 6771–6794. https://doi.org/10.1175/JCLI-D-16-0777.1

PUNGE, H., BEDKA, K., KUNZ, M. and WERNER, A. (2014). A new physically based stochastic event catalog for hail in Europe. *Nat. Hazards* **73** 1625–1645.

RABINER, L. R. and JUANG, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, New Jersey.

RÄDLER, A. T., GROENEMEIJER, P. H., FAUST, E., SAUSEN, R. and PÚČIK, T. (2019). Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. *npj Clim. Atmos. Sci.* **2** 30. https://doi.org/10.1038/s41612-019-0083-7

ROGNA, M., SCHAMEL, G. and WEISSENSTEINER, A. (2019). Modeling the switch from hail insurance to anti-hail nets. *SSRN Electron. J.*. https://doi.org/10.2139/ssrn.3424071

ROHRBECK, C., EASTOE, E. F., FRIGESSI, A. and TAWN, J. A. (2018). Extreme value modelling of water-related insurance claims. *Ann. Appl. Stat.* **12** 246–282. MR3773393 https://doi.org/10.1214/17-AOAS1081

RUE, H., RIEBLER, A., SØRBYE, S. H., ILLIAN, J. B., SIMPSON, D. P. and LINDGREN, F. K. (2017). Bayesian computing with INLA: A review. *Annu. Rev. Stat. Appl.* **4** 395–421.

SCHMID, T., PORTMANN, R., VILLIGER, L., SCHRÖER, K. and BRESCH, D. N. (2024). An open-source radar-based hail damage model for buildings and cars. *Nat. Hazards Earth Syst. Sci.* **24** 847–872. https://doi.org/10.5194/nhess-24-847-2024

SCHMIDBERGER, M. (2018). Hagelgefährdung und Hagelrisiko in Deutschland basierend auf einer Kombination von Radardaten und Versicherungsdaten PhD Thesis Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/KSP/1000086012

SCHUSTER, S. S., BLONG, R. J. and MCANENEY, K. J. (2006). Relationship between radar-derived hail kinetic energy and damage to insured buildings for severe hailstorms in eastern Australia. *Atmos. Res.* **81** 215–235. https://doi.org/10.1016/j.atmosres.2005.12.003

SHARKEY, P., TAWN, J. A. and BROWN, S. J. (2020). Modelling the spatial extent and severity of extreme European windstorms. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 223–250. MR4098947 https://doi.org/10.1111/rssc.12391

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

SHOOTER, R., TAWN, J., 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

TER BRAAK, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Stat. Comput.* **16** 239–249. MR2242236 https://doi.org/10.1007/s11222-006-8769-1

TER BRAAK, C. J. F. and VRUGT, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Stat. Comput.* **18** 435–446. MR2461887 https://doi.org/10.1007/s11222-008-9104-9

VARTY, Z., TAWN, J. A., ATKINSON, P. M. and BIERMAN, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. arXiv:2102.00884.

WARREN, R. A., RAMSAY, H. A., SIEMS, S. T., MANTON, M. J., PETER, J. R., PROTAT, A. and PILLALAMARRI, A. (2020). Radar-based climatology of damaging hailstorms in Brisbane and Sydney, Australia. *Q. J. R. Meteorol. Soc.* **146** 505–530. https://doi.org/10.1002/qj.3693

ZAR, J. H. (2005). Spearman rank correlation. In *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds.) Wiley, New York. https://doi.org/10.1002/0470011815.b2a15150

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3109–3129
https://doi.org/10.1214/24-AOAS1926

MULTIPLE CHANGE POINT DETECTION IN FUNCTIONAL DATA WITH APPLICATIONS TO BIOMECHANICAL FATIGUE DATA

BY PATRICK BASTIAN[1,a], RUPSA BASU[2,c] AND HOLGER DETTE[1,b]

[1]*Faculty of Mathematics, Ruhr-Universität Bochum,* [a]*patrick.bastian@rub.de,* [b]*holger.dette@rub.de*
[2]*Institute of Econometrics and Statistics, Universität zu Köln,* [c]*rbasu@uni-koeln.de*

Injuries to the lower extremity joints are often debilitating, particularly for professional athletes. Understanding the onset of stressful conditions on these joints is, therefore, important in order to ensure prevention of injuries as well as individualised training for enhanced athletic performance. We study the biomechanical joint angles from the hip, knee and ankle for runners who are experiencing fatigue. The data is cyclic in nature and densely collected by body-worn sensors, which makes it ideal to work with in the functional data analysis (FDA) framework.

We develop a new method for multiple change point detection for functional data, which improves the state of the art with respect to at least two novel aspects. First, the curves are compared with respect to their maximum absolute deviation, which leads to a better interpretation of local changes in the functional data compared to classical L^2-approaches. Second, as slight aberrations are to be often expected in a human movement data, our method will not detect arbitrarily small changes but hunts for relevant changes, where maximum absolute deviation between the curves exceeds a specified threshold, say $\Delta > 0$. We recover multiple changes in a long functional time series of biomechanical knee angle data, which are larger than the desired threshold Δ, allowing us to identify changes purely due to fatigue. In this work we analyse data from both controlled indoor as well as from an uncontrolled outdoor (marathon) setting.

REFERENCES

APTE, S., PRIGENT, G., STÖGGL, T., MARTÍNEZ, A., SNYDER, C., GREMEAUX-BADER, V. and AMINIAN, K. (2021). Biomechanical response of the lower extremity to running-induced acute fatigue: A systematic review. *Front. Physiol.* **12** 646042.

ASTON, J. A. D. and KIRCH, C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.* **6** 1906–1948. MR3058688 https://doi.org/10.1214/12-AOAS565

AUE, A., GABRYS, R., HORVÁTH, L. and KOKOSZKA, P. (2009). Estimation of a change-point in the mean function of functional data. *J. Multivariate Anal.* **100** 2254–2269. MR2560367 https://doi.org/10.1016/j.jmva.2009.04.001

AUE, A. and KIRCH, C. (2024). The state of cumulative sum sequential changepoint testing 70 years after Page. *Biometrika* **111** 367–391. MR4745572 https://doi.org/10.1093/biomet/asad079

AUE, A., RICE, G. and SÖNMEZ, O. (2018). Detecting and dating structural breaks in functional data without dimension reduction. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 509–529. MR3798876 https://doi.org/10.1111/rssb.12257

BARANOWSKI, R., CHEN, Y. and FRYZLEWICZ, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 649–672. MR3961502 https://doi.org/10.1111/rssb.12322

BASTIAN, P., BASU, R. and DETTE, H. (2024). Supplement to "Multiple change point detection in functional data with applications to biomechanical fatigue data." https://doi.org/10.1214/24-AOAS1926SUPPA, https://doi.org/10.1214/24-AOAS1926SUPPB

BERKES, I., GABRYS, R., HORVÁTH, L. and KOKOSZKA, P. (2009). Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 927–946. MR2750251 https://doi.org/10.1111/j.1467-9868.2009.00713.x

Key words and phrases. Multiple change detection, relevant changes, functional data, biomechanical joint angles, human gait analysis.

BUCCHIA, B. and WENDLER, M. (2017). Change-point detection and bootstrap for Hilbert space valued random fields. *J. Multivariate Anal.* **155** 344–368. MR3607900 https://doi.org/10.1016/j.jmva.2017.01.007

CHEN, H. (2019). Sequential change-point detection based on nearest neighbors. *Ann. Statist.* **47** 1381–1407. MR3911116 https://doi.org/10.1214/18-AOS1718

CHIOU, J.-M., CHEN, Y.-T. and HSING, T. (2019). Identifying multiple changes for a functional data sequence with application to freeway traffic segmentation. *Ann. Appl. Stat.* **13** 1430–1463. MR4019145 https://doi.org/10.1214/19-AOAS1242

CHO, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10** 2000–2038. MR3522667 https://doi.org/10.1214/16-EJS1155

CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 475–507. MR3310536 https://doi.org/10.1111/rssb.12079

DETTE, H., ECKLE, T. and VETTER, M. (2020). Multiscale change point detection for dependent data. *Scand. J. Stat.* **47** 1243–1274. MR4178193 https://doi.org/10.1111/sjos.12465

DETTE, H., KOKOT, K. and AUE, A. (2020). Functional data analysis in the Banach space of continuous functions. *Ann. Statist.* **48** 1168–1192. MR4102692 https://doi.org/10.1214/19-AOS1842

DETTE, H., KOKOT, K. and VOLGUSHEV, S. (2020). Testing relevant hypotheses in functional time series via self-normalization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 629–660. MR4112779 https://doi.org/10.1111/rssb.12370

DETTE, H. and KUTTA, T. (2021). Detecting structural breaks in eigensystems of functional time series. *Electron. J. Stat.* **15** 944–983. MR4255289 https://doi.org/10.1214/20-ejs1796

EICHINGER, B. and KIRCH, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* **24** 526–564. MR3706768 https://doi.org/10.3150/16-BEJ887

FERNIQUE, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École D'Été de Probabilités de Saint-Flour, IV*-1974. *Lecture Notes in Math.* **480** 1–96. Springer, Berlin. MR0413238

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. With 32 discussions by 47 authors and a rejoinder by the authors. MR3210728 https://doi.org/10.1111/rssb.12047

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

HARRIS, T., LI, B. and TUCKER, J. D. (2022). Scalable multiple changepoint detection for functional data sequences. *Environmetrics* **33** Paper No. e2710, 17. MR4393413 https://doi.org/10.1002/env.2710

HORVÁTH, L., KOKOSZKA, P. and RICE, G. (2014). Testing stationarity of functional time series. *J. Econometrics* **179** 66–82. MR3153649 https://doi.org/10.1016/j.jeconom.2013.11.002

HORVÁTH, L., LIU, Z., RICE, G., WANG, S. and ZHAN, Y. (2023). Testing stability in functional event observations with an application to IPO performance. *J. Bus. Econom. Statist.* **41** 1262–1273. MR4650460 https://doi.org/10.1080/07350015.2022.2118127

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418 https://doi.org/10.1080/01621459.2012.737745

KOVÁCS, S., BÜHLMANN, P., LI, H. and MUNK, A. (2023). Seeded binary segmentation: A general methodology for fast and optimal changepoint detection. *Biometrika* **110** 249–256. MR4565454 https://doi.org/10.1093/biomet/asac052

LI, H., GUO, Q. and MUNK, A. (2019). Multiscale change-point segmentation: Beyond step functions. *Electron. J. Stat.* **13** 3254–3296. MR4010980 https://doi.org/10.1214/19-ejs1608

MAAS, E., BIE, J. D., VANFLETEREN, R., HOOGKAMER, W. and VANWANSEELE, B. (2018). Novice runners show greater changes in kinematics with fatigue compared with competitive runners. *Sports Biomech.* **17** 350–360. https://doi.org/10.1080/14763141.2017.1347193

MADRID PADILLA, C. M., WANG, D., ZHAO, Z. and YU, Y. (2022a). Change-point detection for sparse and dense functional data in general dimensions. In *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, eds.) **35** 37121–37133. Curran Associates, Red Hook.

MADRID PADILLA, O. H., YU, Y., WANG, D. and RINALDO, A. (2022b). Optimal nonparametric multivariate change point detection and localization. *IEEE Trans. Inf. Theory* **68** 1922–1944. MR4395506 https://doi.org/10.1109/TIT.2021.3130330

MAIDSTONE, R., HOCKING, T., RIGAILL, G. and FEARNHEAD, P. (2017). On optimal multiple changepoint algorithms for large data. *Stat. Comput.* **27** 519–533. MR3599687 https://doi.org/10.1007/s11222-016-9636-3

RICE, G. and SHANG, H. L. (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *J. Time Series Anal.* **38** 591–609. MR3664648 https://doi.org/10.1111/jtsa.12229

RICE, G. and ZHANG, C. (2022). Consistency of binary segmentation for multiple change-point estimation with functional data. *Statist. Probab. Lett.* **180** Paper No. 109228, 8. MR4318362 https://doi.org/10.1016/j.spl.2021.109228

SCHEPERS, M., GIUBERTI, M., BELLUSCI, G. et al. (2018). Xsens MVN: Consistent tracking of human motion using inertial sensing. Xsens Technol 1. https://doi.org/10.13140/RG.2.2.22099.07205

SHAPIROV, O., TEWES, J. and WENDLER, M. (2016). Sequential block bootstrap in a Hilbert space with application to change point analysis. *Canad. J. Statist.* **44** 300–322. MR3536199 https://doi.org/10.1002/cjs.11293

STOEHR, C., ASTON, J. A. D. and KIRCH, C. (2021). Detecting changes in the covariance structure of functional time series with application to fMRI data. *Econom. Stat.* **18** 44–62. MR4238907 https://doi.org/10.1016/j.ecosta.2020.04.004

STÖHR, C. (2019). Sequential change point procedures based on U-statistics and the detection of covariance changes in functional data. PhD thesis, Dissertation, Magdeburg, Otto-von-Guericke-Universität Magdeburg, 2019.

TUKEY, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* **6** 100–116.

VERZELEN, N., FROMONT, M., LERASLE, M. and REYNAUD-BOURET, P. (2023). Optimal change-point detection and localization. *Ann. Statist.* **51** 1586–1610. MR4658569 https://doi.org/10.1214/23-aos2297

VOSTRIKOVA, L. Y. (1981). Discovery of "discord" in multidimensional random processes. *Dokl. Akad. Nauk SSSR* **259** 270–274. MR0625215

WANG, D., YU, Y. and RINALDO, A. (2020). Univariate mean change point detection: Penalization, CUSUM and optimality. *Electron. J. Stat.* **14** 1917–1961. MR4091859 https://doi.org/10.1214/20-EJS1710

WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. MR3744712 https://doi.org/10.1111/rssb.12243

ZANDBERGEN, M. A., BUURKE, J. H., VELTINK, P. H. and REENALDA, J. (2023a). Quantifying and correcting for speed and stride frequency effects on running mechanics in fatiguing outdoor running. *Front. Sports Act. Living* **5** 1085513.

ZANDBERGEN, M. A., MAROTTA, L., BULTHUIS, R., BUURKE, J. H., VELTINK, P. H. and REENALDA, J. (2023b). Effects of level running-induced fatigue on running kinematics: A systematic review and meta-analysis. *Gait Posture* **99** 60–75.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3130–3145
https://doi.org/10.1214/24-AOAS1927

UTILIZING A CAPTURE–RECAPTURE STRATEGY TO ACCELERATE INFECTIOUS DISEASE SURVEILLANCE

BY LIN GE[a], YUZI ZHANG[b], LANCE WALLER[c] AND ROBERT LYLES[d]

Department of Biostatistics and Bioinformatics, Emory University, [a]*lge_biostat@hotmail.com,* [b]*yuzi.zhang@emory.edu,* [c]*lwaller@emory.edu,* [d]*rlyles@emory.edu*

Monitoring key elements of disease dynamics (e.g., prevalence, case counts) is of great importance in infectious disease prevention and control, as emphasized during the COVID-19 pandemic. To facilitate this effort, we propose a new capture–recapture (CRC) analysis strategy that adjusts for misclassification stemming from the use of easily administered but imperfect diagnostic test kits, such as rapid antigen test-kits or saliva tests. Our method is based on a recently proposed "anchor stream" design, whereby an existing voluntary surveillance data stream is augmented by a smaller and judiciously drawn random sample. It incorporates manufacturer-specified sensitivity and specificity parameters to account for imperfect diagnostic results in one or both data streams. For inference to accompany case count estimation, we improve upon traditional Wald-type confidence intervals by developing an adapted Bayesian credible interval for the CRC estimator that yields favorable frequentist coverage properties. When feasible, the proposed design and analytic strategy provides a more efficient solution than traditional CRC methods or random sampling-based bias-corrected estimation to monitor disease prevalence while accounting for misclassification. We demonstrate the benefits of this approach through simulation studies and a numerical example that underscore its potential utility in practice for economical disease monitoring among a registered closed population.

REFERENCES

ADAMS, G. (2020). A beginner's guide to RT-PCR, qPCR and RT-qPCR. *Biochemist* **42** 48–53.

AGRESTI, A. (1994). Simple capture–recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50** 494–500. MR0556485

AGRESTI, A. and COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Statist*. https://doi.org/10.1080/00031305.1998.10480550

BAILLARGEON, S. and RIVEST, L.-P. (2007). Rcapture: Loglinear models for capture–recapture in R. *J. Stat. Softw.* **19** 1–31.

BLYTH, C. R. and STILL, H. A. (1983). Binomial confidence intervals. *J. Amer. Statist. Assoc*. https://doi.org/10.1080/01621459.1983.10477938

BRENNER, H. (1995). Use and limitations of the capture–recapture method in disease monitoring with two dependent sources. *Epidemiology* **6** 42–48. https://doi.org/10.1097/00001648-199501000-00009

BRENNER, H. (1996). Effects of misdiagnoses on disease monitoring with capture–recapture methods. *J. Clin. Epidemiol*. **49** 1303–1307. https://doi.org/10.1016/0895-4356(95)00026-7

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci*. **16** 101–133.

CARVALHO, C., ALBA, S., HARRIS, R., ABUBAKAR, I., VAN HEST, R., CORREIA, A. M., GONÇALVES, G. and DUARTE, R. (2020). Completeness of TB notification in Portugal, 2015: An inventory and capture–recapture study. *Int. J. Tuberc. Lung Dis*. https://doi.org/10.5588/IJTLD.20.0094

CASATI, B., VERDI, J. P., HEMPELMANN, A., KITTEL, M., KLAEBISCH, A. G., MEISTER, B., WELKER, S., ASTHANA, S., DI GIORGIO, S. et al. (2022). Rapid, adaptable and sensitive Cas13-based COVID-19 diagnostics using ADESSO. *Nat. Commun*. **13** 3308.

CHAO, A., PAN, H.-Y. and CHIANG, S.-C. (2008). The Petersen–Lincoln estimator and its extension to estimate the size of a shared population. *Biom. J.* **50** 957–970.

Key words and phrases. Credible interval, misclassification, nonrepresentative sampling, sensitivity, specificity.

CHAPMAN, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological simple censuses. Univ. California Publications in Statistics.

CHATTERJEE, K. and MUKHERJEE, D. (2016). On the estimation of homogeneous population size from a complex dual-record system. *J. Stat. Comput. Simul.* **86** 3562–3581.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

CORMACK, R. M. (1999). Problems with using capture–recapture in epidemiology: An example of a measles epidemic. *J. Clin. Epidemiol.* **52** 909–914. https://doi.org/10.1016/s0895-4356(99)00058-x

DUNBAR, R., VAN HEST, R., LAWRENCE, K., VERVER, S., ENARSON, D. A., LOMBARD, C., BEYERS, N. and BARNES, J. M. (2011). Capture–recapture to estimate completeness of tuberculosis surveillance in two communities in South Africa. *Int. J. Tuberc. Lung Dis.* **15** 1038–1043. https://doi.org/10.5588/ijtld.10.0695

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 591–603.

GASTWIRTH, J. L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statist. Sci.* **2** 213–222.

GE, L., ZHANG, Y., WALLER, L. and LYLES, R. (2024). Supplement to "Utilizing a Capture–Recapture Strategy to Accelerate Infectious Disease Surveillance." https://doi.org/10.1214/24-AOAS1927SUPPA, https://doi.org/10.1214/24-AOAS1927SUPPB

GE, L., ZHANG, Y., WALLER, L. A. and LYLES, R. H. (2023a). Enhanced inference for finite population sampling-based prevalence estimation with misclassification errors. *Amer. Statist.* https://doi.org/10.1080/00031305.2023.2250401

GE, L., ZHANG, Y., WARD, K. C., LASH, T. L., WALLER, L. A. and LYLES, R. H. (2023b). Tailoring capture–recapture methods to estimate registry-based case counts based on error-prone diagnostic signals. *Stat. Med.* **42** 2928–2943. https://doi.org/10.1002/sim.9759

GHOSH, B. K. (1979b). A comparison of some approximate confidence intervals for the binomial parameter. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.1979.10481051

GLASZIOU, P., IRWIG, B. and DEEKS, J. J. (2008). When should a new test become the current reference standard? *Ann. Intern. Med.* **149** 816–821. MR0686755

HOOK, E. B. and REGAL, R. R. (1995). Capture–recapture methods in epidemiology: Methods and limitations. *Epidemiol. Rev.* **17** 243–264.

HOPKINS, R. S. (2005). Design and operation of state and local infectious disease surveillance systems. *J. Public Health Manag. Pract.* https://doi.org/10.1097/00124784-200505000-00002

JONES, H. E., HICKMAN, M., WELTON, N. J., ANGELIS, D. D., HARRIS, R. J. and ADES, A. E. (2014). Recapture or precapture? Fallibility of standard capture–recapture methods in the presence of referrals between sources. *Amer. J. Epidemiol.* **179** 1383–1393. https://doi.org/10.1093/aje/kwu056

LEVY, P. S. and KASS, E. H. (1970). A three-population model for sequential screening for bacteriuria. *Amer. J. Epidemiol.* **91** 148–154. https://doi.org/10.1093/oxfordjournals.aje.a121122

LINCOLN, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. U.S. Dept. Agriculture.

LYLES, R. H., ZHANG, Y., GE, L., ENGLAND, C., WARD, K., LASH, T. L. and WALLER, L. A. (2022). Using capture–recapture methodology to enhance precision of representative sampling-based case count estimates. *J. Surv. Stat. Methodol.*

LYLES, R. H., ZHANG, Y., GE, L. and WALLER, L. A. (2023). A design and analytic strategy for monitoring disease positivity and case characteristics in accessible closed populations. *Amer. J. Epidemiol.* https://doi.org/10.1093/aje/kwad177

MATHESON, N. J., WARNE, B., WEEKES, M. P. and MAXWELL, P. H. (2021). Mass testing of university students for covid-19. *BMJ*. https://doi.org/10.1136/bmj.n2388

MENNI, C., VALDES, A. M., FREIDIN, M. B., SUDRE, C. H., NGUYEN, L. H., DREW, D. A., GANESH, S., VARSAVSKY, T., CARDOSO, M. J. et al. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat. Med.* https://doi.org/10.1038/s41591-020-0916-2

MURAKAMI, M., SATO, H., IRIE, T., KAMO, M., NAITO, W., YASUTAKA, T. and IMOTO, S. (2023). Sensitivity of rapid antigen tests for COVID-19 during the Omicron variant outbreak among players and staff members of the Japan Professional Football League and clubs: A retrospective observational study. *BMJ Open* **13** e067591. https://doi.org/10.1136/bmjopen-2022-067591

PEREZ DUQUE, M., HANSEN, L., ANTUNES, D. and SÁ MACHADO, R. (2020). Capture–recapture study to estimate the true incidence of tuberculosis in Portugal, 2018. *Eur. J. Public Health* **30** ckaa165–795.

PETERSEN, C. G. J. (1986). The yearly immigration of young plaice into the Limfjord from the German Sea. *Rep. Dan. Biol. Stn.* 1985 **6** 1–48.

POOROLAJAL, J., MOHAMMADI, Y. and FARZINARA, F. (2017). Using the capture–recapture method to estimate the human immunodeficiency virus-positive population. *Epidemiol. Health* **39** e2017042. https://doi.org/10.4178/epih.e2017042

RAMOS, P. L., SOUSA, I., SANTANA, R., MORGAN, W. H., GORDON, K., CREWE, J., ROCHA-SOUSA, A. and MACEDO, A. F. (2020). A review of capture–recapture methods and its possibilities in ophthalmology and vision sciences. *Ophthalmic Epidemiol.* **27** 310–324. https://doi.org/10.1080/09286586.2020.1749286

RENNERT, L. and MCMAHAN, C. (2022). Risk of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Reinfection in a University Student Population. Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciab454

RENNERT, L., MCMAHAN, C., KALBAUGH, C. A., YANG, Y., LUMSDEN, B., DEAN, D., PEKAREK, L. and COLENDA, C. C. (2021). Surveillance-based informative testing for detection and containment of SARS-CoV-2 outbreaks on a public university campus: An observational and modelling study. *Lancet Child Adolesc. Health*. https://doi.org/10.1016/S2352-4642(21)00060-2

ROGAN, W. J. and GLADEN, B. (1978). Estimating prevalence from the results of a screening test. *Amer. J. Epidemiol.* **107** 71–76. https://doi.org/10.1093/oxfordjournals.aje.a112510

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

SCHULTES, O., CLARKE, V., PALTIEL, A. D., CARTTER, M., SOSA, L. and CRAWFORD, F. W. (2021). COVID-19 testing and case rates and social contact among residential college students in Connecticut during the 2020-2021 academic year. *JAMA Netw. Open*. https://doi.org/10.1001/jamanetworkopen.2021.40602

SEBER, G. A. F. et al. (1982). *The Estimation of Animal Abundance and Related Parameters*. Blackburn press, Caldwell, NJ.

SOH, B., LEE, W., KENCH, P., REED, W., MCENTEE, M., POULOS, A. and BRENNAN, P. (2012a). Assessing reader performance in radiology, an imperfect science: Lessons from breast screening. *Clin. Radiol.* **67** 623–628.

SOH, S. E., COOK, A. R., CHEN, M. I. C., LEE, V. J., CUTTER, J. L., CHOW, V. T. K., TEE, N. W. S., LIN, R. T. P., LIM, W. Y. et al. (2012b). Teacher led school-based surveillance can allow accurate tracking of emerging infectious diseases—evidence from serial cross-sectional surveys of febrile respiratory illness during the H1N1 2009 influenza pandemic in Singapore. *BMC Infect. Dis.* https://doi.org/10.1186/1471-2334-12-336

SUYAMA, J., SZTAJNKRYCER, M., LINDSELL, C., OTTEN, E. J., DANIELS, J. M. and KRESSEL, A. B. (2003). Surveillance of infectious disease occurrences in the community: An analysis of symptom presentation in the emergency department. *Acad. Emerg. Med.* https://doi.org/10.1197/aemj.10.7.753

VANDER SCHAAF, N. A., FUND, A. J., MUNNICH, B. V., ZASTROW, A. L., FUND, E. E., SENTI, T. L., LYNN, A. F., KANE, J. J., LOVE, J. L. et al. (2021). Routine, cost-effective SARS-CoV-2 surveillance testing using pooled saliva limits viral spread on a residential college campus. *Microbiol. Spectrum* 01089–21. https://doi.org/10.1128/spectrum

WALTER, S. D., MACASKILL, P., LORD, S. J. and IRWIG, L. (2012). Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. *Stat. Med.* **31** 1129–1138. https://doi.org/10.1002/sim.4444

WEINBERG, M., WATERMAN, S., ALVAREZ LUCAS, C., CARRION FALCON, V., KURI MORALES, P., ANAYA LOPEZ, L., PETER, C., ESCOBAR GUTIÉRREZ, A., RAMIREZ GONZALEZ, E. et al. (2003). The U.S.-Mexico Border Infectious Disease Surveillance project: Establishing binational border surveillance. Emerging Infectious Diseases. https://doi.org/10.3201/eid0901.020047

WU, C., CHANG, H.-G., MCNUTT, L.-A. and SMITH, P. (2005). Estimating the mortality rate of hepatitis C using multiple data sources. *Epidemiol. Infect.* **133** 121–125.

ZHANG, B. and SMALL, D. S. (2020). Number of healthcare workers who have died of COVID-19. *Epidemiology* **31** e46. https://doi.org/10.1097/EDE.0000000000001229

ZHANG, Y., CHEN, J., GE, L., WILLIAMSON, J. M., WALLER, L. A. and LYLES, R. H. (2023a). Sensitivity and uncertainty analysis for two-stream capture–recapture methods in disease surveillance. *Epidemiology* **34** 601–610.

ZHANG, Y., GE, L., WALLER, L. A. and LYLES, R. H. (2023b). On some pitfalls of the log-linear modeling framework for capture–recapture studies in disease surveillance. *Epidemiol. Methods* **12** 20230019.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3146–3164
https://doi.org/10.1214/24-AOAS1928

A BAYESIAN MODEL OF UNDERREPORTING FOR SEXUAL ASSAULT ON COLLEGE CAMPUSES

BY CASEY BRADSHAW[a] AND DAVID M. BLEI[b]

Department of Statistics, Columbia University, [a]*cb3431@columbia.edu,* [b]*david.blei@columbia.edu*

In an effort to quantify and combat sexual assault, U.S. colleges and universities are required to disclose the number of reported sexual assaults on their campuses each year. However, many instances of sexual assault are never reported to authorities, and consequently, the number of reported assaults does not fully reflect the true total number of assaults that occurred; the reported values could arise from many combinations of reporting rate and true incidence. In this paper we estimate these underlying quantities via a hierarchical Bayesian model of the reported number of assaults. We use informative priors, based on national crime statistics, to act as a tiebreaker to help distinguish between reporting rates and incidence. We outline a Hamiltonian Monte Carlo (HMC) sampling scheme for posterior inference regarding reporting rates and assault incidence at each school and apply this method to campus sexual assault data from 2014–2019. Results suggest an increasing trend in reporting rates for the overall college population during this time. However, the extent of underreporting varies widely across schools. That variation has implications for how individual schools should interpret their reported crime statistics.

REFERENCES

BAILEY, T. C., CARVALHO, M. S., LAPA, T. M., SOUZA, W. V. and BREWER, M. J. (2005). Modeling of underdetection of cases in disease surveillance. *Ann. Epidemiol.* **15** 335–343. https://doi.org/10.1016/j.annepidem.2004.09.013

BETTENCOURT, L. M. A., LOBO, J., HELBING, D., KÜHNERT, C. and WEST, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. USA* **104** 7301–7306.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. With discussion. MR0603745 https://doi.org/10.2307/2982063

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

BRADSHAW, C. and BLEI, D. M. (2024). Supplement to "A Bayesian Model of Underreporting for Sexual Assault on College Campuses." https://doi.org/10.1214/24-AOAS1928SUPPA, https://doi.org/10.1214/24-AOAS1928SUPPB

CHANG, Y., CHOI, S. B., LEE, J. and JIN, W. C. (2018). Population size vs. Number of crimes: Is the relationship superlinear? *Int. J. Inf. Syst. Soc. Change* **9** 26–39.

DE OLIVEIRA, G. L., LOSCHI, R. H. and ASSUNÇÃO, R. M. (2017). A random-censoring Poisson model for underreported data. *Stat. Med.* **36** 4873–4892. MR3734480 https://doi.org/10.1002/sim.7456

DE OLIVEIRA, G. L., OLIVEIRA, J. F., PESCARINI, J. M., ANDRADE, R. F. S., NERY, J. S., ICHIHARA, M. Y., SMEETH, L., BRICKLEY, E. B., BARRETO, M. L. et al. (2021). Estimating underreporting of leprosy in Brazil using a Bayesian approach. *PLoS Negl. Trop. Dis.* **15**.

DVORZAK, M. and WAGNER, H. (2016). Sparse Bayesian modelling of underreported count data. *Stat. Model.* **16** 24–46. MR3457686 https://doi.org/10.1177/1471082X15588398

FADER, P. S. and HARDIE, B. G. (2000). A note on modelling underreported Poisson counts. *J. Appl. Stat.* **27** 953–964.

FERNÁNDEZ-FONTELO, A., CABAÑA, A., JOE, H., PUIG, P. and MORIÑA, D. (2019). Untangling serially dependent underreported count data for gender-based violence. *Stat. Med.* **38** 4404–4422. MR4002439 https://doi.org/10.1002/sim.8306

Key words and phrases. Underreporting, Bayesian hierarchical model, count data.

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

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

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

GUTTMAN, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *J. Roy. Statist. Soc. Ser. B* **29** 83–100. MR0216699

JAYNES, E. T. (1968). Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.* **4** 227–241.

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

KUMARA, S. and CHIN, H. C. (2005). Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. *Transp. Res. Rec.* **1908** 46–50.

LAURITSEN, J. L., GATEWOOD OWENS, J., PLANTY, M., RAND, M. R. and TRUMAN, J. L. (2012). Methods for Counting High-Frequency Repeat Victimizations in the National Crime Victimization Survey. Bureau of Justice Statistics NCJ 237308.

LETOUZEY, F., DENIS, F. and GILLERON, R. (2000). Learning from positive and unlabeled examples. In *Algorithmic Learning Theory* (*Sydney*, 2000). *Lecture Notes in Computer Science* **1968** 71–85. Springer, Berlin. MR1851968 https://doi.org/10.1007/3-540-40992-0_6

LI, J. and HUGGINS, J. H. (2022). Calibrated model criticism using split predictive checks. arXiv preprint. Available at arXiv:2203.15897.

MA, J. and LI, Z. (2010). Bayesian modeling of frequency-severity indeterminacy with an application to traffic crashes on two-lane highways. In *ICCTP* 2010: *Integrated Transportation Systems*: *Green*, *Intelligent*, *Reliable* 1022–1033.

MORAN, G. E., BLEI, D. M. and RANGANATH, R. (2019). Population predictive checks. ArXiv preprint. Available at arXiv:1908.00882.

MORENO, E. and GIRÓN, J. (1998). Estimating with incomplete count data: A Bayesian approach. *J. Statist. Plann. Inference* **66** 147–159. MR1617002 https://doi.org/10.1016/S0378-3758(97)00073-6

MORGAN, R. E. and KENA, G. (2018). Criminal Victimization, 2016: Revised. Bureau of Justice Statistics NCJ 252121.

MORGAN, R. E. and THOMPSON, A. (2020). Criminal Victimization, 2019. Bureau of Justice Statistics NCJ 255113.

MORGAN, R. E. and TRUMAN, J. L. (2018). Criminal Victimization, 2017. Bureau of Justice Statistics NCJ 252472.

OBAMA, B. (2014). Memorandum—Establishing a White House Task Force to Protect Students from Sexual Assault. The White House Office of the Press Secretary. 22 Jan 2014. https://obamawhitehouse.archives.gov/the-press-office/2014/01/22/memorandum-establishing-white-house-task-force-protect-students-sexual-a (accessed: 17 Aug 2023).

POWERS, S., GERLACH, R. and STAMEY, J. (2010). Bayesian variable selection for Poisson regression with underreported responses. *Comput. Statist. Data Anal.* **54** 3289–3299. MR2727752 https://doi.org/10.1016/j.csda.2010.04.003

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

SABINA, C. and HO, L. Y. (2014). Campus and college victim responses to sexual assault and dating violence: Disclosure, service utilization, and service provision. *Trauma Violence Abuse* **15** 201–226. https://doi.org/10.1177/1524838014521322

SCHMERTMANN, C. P. and GONZAGA, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography* **55** 1363–1388. https://doi.org/10.1007/s13524-018-0695-2

SHANMUGAM, D. and PIERSON, E. (2021). Quantifying inequality in underreported medical conditions. ArXiv preprint. Available at arXiv:2110.04133.

SINOZICH, S. and LANGTON, L. (2014). Rape and sexual assault victimization among college-age females, 1995-2013. Bureau of Justice Statistics NCJ 248471.

SOMANADER, T. (2014). President Obama Launches the "It's On Us" Campaign to End Sexual Assault on Campus. The White House Blog 19 Sept 2014. https://obamawhitehouse.archives.gov/blog/2014/09/19/president-obama-launches-its-us-campaign-end-sexual-assault-campus (accessed: 17 Aug 2023).

STAN DEVELOPMENT TEAM (2023). RStan: the R interface to Stan. R package version 2.21.8.

STONER, O., ECONOMOU, T. and DRUMMOND MARQUES DA SILVA, G. (2019). A hierarchical framework for correcting under-reporting in count data. *J. Amer. Statist. Assoc.* **114** 1481–1492. MR4047275 https://doi.org/10.1080/01621459.2019.1573732

TRUMAN, J. L. and LANGTON, L. (2014). Criminal Victimization, 2013 (Revised). Bureau of Justice Statistics NCJ 247648.

TRUMAN, J. L. and LANGTON, L. (2015). Criminal Victimization, 2014. Bureau of Justice Statistics NCJ 248973.

TRUMAN, J. L. and MORGAN, R. E. (2016). Criminal Victimization, 2015. Bureau of Justice Statistics NCJ 250180.

WINKELMANN, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empir. Econ*. **21** 575–587.

WU, K., DAHLEM, D., HANE, C., HALPERIN, E. and ZOU, J. (2023). Collecting data when missingness is unknown: A method for improving model performance given under-reporting in patient populations. In *Conference on Health, Inference and Learning* 229–242. PMLR.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3165–3184
https://doi.org/10.1214/24-AOAS1930

DYNAMIC TOPIC LANGUAGE MODEL ON HETEROGENEOUS CHILDREN'S MENTAL HEALTH CLINICAL NOTES

BY HANWEN YE[1,a], TATIANA MORENO[2,c], ADRIANNE ALPERN[2,d], LOUIS EHWERHEMUEPHA[2,e] AND ANNIE QU[1,b]

[1]*Department of Statistics, University of California,* [a]*hanweny@uci.edu,* [b]*aqu2@uci.edu*
[2]*Children's Hospital of Orange County,* [c]*Tatiana.Moreno@choc.org,* [d]*AAlpern@choc.org,* [e]*LEhwerhemuepha@choc.org*

Mental health diseases which affect children's lives and well-beings have received increased attention since the COVID-19 pandemic. Analyzing psychiatric clinical notes with topic models is critical to evaluating children's mental status over time. However, few topic models are built for longitudinal settings, and most existing approaches fail to capture temporal trajectories for each document. To address these challenges, we develop a dynamic topic model with consistent topics and individualized temporal dependencies on the evolving document metadata. Our model preserves the semantic meaning of discovered topics over time and incorporates heterogeneity among documents. In particular, when documents can be categorized, we propose a classifier-free approach to maximize topic heterogeneity across different document groups. We also present an efficient variational optimization procedure adapted for the multistage longitudinal setting. In this case study, we apply our method to the psychiatric clinical notes from a large tertiary pediatric hospital in Southern California and achieve a 38% increase in the overall coherence of extracted topics. Our real data analysis reveals that children tend to express more negative emotions during state shutdowns and more positive when schools reopen. Furthermore, it suggests that sexual and gender minority (SGM) children display more pronounced reactions to major COVID-19 events and a greater sensitivity to vaccine-related news than non-SGM children. This study examines children's mental health progression during the pandemic and offers clinicians valuable insights to recognize disparities in children's mental health related to their sexual and gender identities.

REFERENCES

ADZRAGO, D., ORMISTON, C. K., SULLEY, S. and WILLIAMS, F. (2023). Associations between the self-reported likelihood of receiving the COVID-19 vaccine, likelihood of contracting COVID-19, discrimination, and anxiety/depression by sexual orientation. *Vaccines* **11** 582.

AFIFI, M. (2007). Gender differences in mental health. *Singapore Medical Journal* **48** 385.

BARRY, T. R. (2014). The midlife in the United States (MIDUS) series: A national longitudinal study of health and well-being. *Open Health Data* **2**.

BLEI, D. and LAFFERTY, J. (2006a). Correlated topic models. *Adv. Neural Inf. Process. Syst.* **18** 147.

BLEI, D. M. and LAFFERTY, J. D. (2006b). Dynamic topic models. In *Proceedings of the* 23*rd International Conference on Machine Learning* 113–120.

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

BOYD, A., GOLDING, J., MACLEOD, J., LAWLOR, D. A., FRASER, A., HENDERSON, J., MOLLOY, L., NESS, A., RING, S. et al. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the avon longitudinal study of parents and children. *Int. J. Epidemiol.* **42** 111–127.

CARD, D., TAN, C. and SMITH, N. A. (2017). Neural models for documents with metadata. Preprint. Available at arXiv:1705.09296.

CASALE, F. P., DALCA, A., SAGLIETTI, L., LISTGARTEN, J. and FUSI, N. (2018). Gaussian process prior variational autoencoders. *Adv. Neural Inf. Process. Syst.* **31**.

Key words and phrases. Classifier-free, multistage topic language models, sexual and gender identity, time-consistent topics, variational inference.

CIECHANOWSKI, K., JEMIELNIAK, D. and SILCZUK, A. (2023). Public interests in mental health topics in COVID-19: Evidence from Wikipedia searches. *Adv. Mental Health* 1–22.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2023). CDC Museum COVID-19 Timeline.

FORTUIN, V., BARANCHUK, D., RÄTSCH, G. and MANDT, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *International Conference on Artificial Intelligence and Statistics* 1651–1661. PMLR.

GATES, G. J. (2014). LGBT demographics: Comparisons among population-based surveys.

GULRAJANI, I., KUMAR, K., AHMED, F., TAIGA, A. A., VISIN, F., VAZQUEZ, D. and COURVILLE, A. (2016). Pixelvae: A latent variable model for natural images. Preprint. Available at arXiv:1611.05013.

GUPTA, P., CHAUDHARY, Y., BUETTNER, F. and SCHÜTZE, H. (2019). Document informed neural autoregressive topic models with distributional prior. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 6505–6512.

HU, X., WANG, R., ZHOU, D. and XIONG, Y. (2020). Neural topic modeling with cycle-consistent adversarial training. Preprint. Available at arXiv:2009.13971.

KARIM, S., CHOUKAS-BRADLEY, S., RADOVIC, A., ROBERTS, S. R., MAHEUX, A. J. and ESCOBAR-VIERA, C. G. (2022). Support over social media among socially isolated sexual and gender minority youth in rural US during the COVID-19 pandemic: Opportunities for intervention research. *Int. J. Environ. Res. Public Health* **19** 15611.

LAROCHELLE, H. and LAULY, S. (2012). A neural autoregressive topic model. *Adv. Neural Inf. Process. Syst.* **25**.

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791. https://doi.org/10.1038/44565

LI, X., OUYANG, J. and ZHOU, X. (2015). Supervised topic models for multi-label classification. *Neurocomputing* **149** 811–819.

LI, Y., ZHU, R., QU, A., YE, H. and SUN, Z. (2021). Topic modeling on triage notes with semiorthogonal non-negative matrix factorization. *J. Amer. Statist. Assoc.* **116** 1609–1624. MR4353700 https://doi.org/10.1080/01621459.2020.1862667

LIN, T., HU, Z. and GUO, X. (2019). Sparsemax and relaxed Wasserstein for topic sparsity. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* 141–149.

MARSHAL, M. P., DIETZ, L. J., FRIEDMAN, M. S., STALL, R., SMITH, H. A., MCGINLEY, J., THOMA, B. C., MURRAY, P. J., D'AUGELLI, A. R. et al. (2011). Suicidality and depression disparities between sexual minority and heterosexual youth: A meta-analytic review. *J. Adolesc. Health* **49** 115–123. https://doi.org/10.1016/j.jadohealth.2011.02.005

MCAULIFFE, J. and BLEI, D. (2007). Supervised topic models. *Adv. Neural Inf. Process. Syst.* **20**.

MCGEOUGH, B. L. and STERZING, P. R. (2018). A systematic review of family victimization experiences among sexual minority youth. *J. Prim. Prev.* **39** 491–528. https://doi.org/10.1007/s10935-018-0523-x

MCGREGOR, K., WILLIAMS, C. R., BOTTA, A., MANDEL, F. and GENTILE, J. (2023). Providing essential gender-affirming telehealth services to transgender youth during COVID-19: A service review. *J. Telemed. Telecare* **29** 147–152. https://doi.org/10.1177/1357633X221095785

MIAO, Y., YU, L. and BLUNSOM, P. (2016). Neural variational inference for text processing. In *International Conference on Machine Learning* 1727–1736. PMLR.

NEWMAN, D., NOH, Y., TALLEY, E., KARIMI, S. and BALDWIN, T. (2010). Evaluating topic models for digital libraries. In *Proceedings of the* 10*th Annual Joint Conference on Digital Libraries* 215–224.

NATIONAL INSTITUTE OF MENTAL HEALTH (2021). Mental health topics. From https://www.nimh.nih.gov/health/topics.

PAATERO, P. and TAPPER, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5** 111–126.

PENNINX, B. W., BEEKMAN, A. T., SMIT, J. H., ZITMAN, F. G., NOLEN, W. A., SPINHOVEN, P., CUIJPERS, P., DE JONG, P. J., VAN MARWIJK, H. W. et al. (2008). The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17** 121–140.

PHARR, J. R., TERRY, E., WADE, A., HABOUSH-DELOYE, A., MARQUEZ, E., HEALTH, N. M. and COALITION, E. (2022). Impact of COVID-19 on sexual and gender minority communities: Focus group discussions. *Int. J. Environ. Res. Public Health* **20** 50.

PLÖDERL, M. and TREMBLAY, P. (2015). Mental health of sexual minorities. A systematic review. *Int. Rev. Psychiatry* **27** 367–385. https://doi.org/10.3109/09540261.2015.1083949

RAMCHANDRAN, S., TIKHONOV, G., KUJANPÄÄ, K., KOSKINEN, M. and LÄHDESMÄKI, H. (2021). Longitudinal variational autoencoder. In *International Conference on Artificial Intelligence and Statistics* 3898–3906. PMLR.

RAVENS-SIEBERER, U., KAMAN, A., ERHART, M., DEVINE, J., SCHLACK, R. and OTTO, C. (2022). Impact of the COVID-19 pandemic on quality of life and mental health in children and adolescents in Germany. *Eur. Child Adolesc. Psychiatry* **31** 879–889.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

ROBERTS, M. E., STEWART, B. M., TINGLEY, D., LUCAS, C., LEDER-LUIS, J., GADARIAN, S. K., ALBERTSON, B. and RAND, D. G. (2014). Structural topic models for open-ended survey responses. *Amer. J. Polit. Sci.* **58** 1064–1082.

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.

RONALD, W., CAROL, D., ELSIE, J., LELA, R., SATVINDER, D. and TARA, W. (2010). Evolving definitions of mental illness and wellness. *Prev. Chronic Dis.* **7** 2.

ROSENFIELD, S. and MOUZON, D. (2013). Gender and mental health. *Handbook of the Sociology of Mental Health* 277–296.

RUSSELL, S. T. and FISH, J. N. (2016). Mental health in lesbian, gay, bisexual, and transgender (LGBT) youth. *Annu. Rev. Clin. Psychol.* **12** 465–487. https://doi.org/10.1146/annurev-clinpsy-021815-093153

SALERNO, J. P., DEVADAS, J., PEASE, M., NKETIA, B. and FISH, J. N. (2020). Sexual and gender minority stress amid the COVID-19 pandemic: Implications for LGBTQ young persons' mental health and well-being. *Public Health Rep.* **135** 721–727. https://doi.org/10.1177/0033354920954511

SCOTT, W. A. (1958). Research definitions of mental health and mental illness. *Psychol. Bull.* **55** 29.

SGARRO, A. (1981). Informational divergence and the dissimilarity of probability distributions. *Calcolo* **18** 293–302. MR0647828 https://doi.org/10.1007/BF02576360

SHARIFIAN-ATTAR, V., DE, S., JABBARI, S., LI, J., MOSS, H. and JOHNSON, J. (2022). Analysing longitudinal social science questionnaires: Topic modelling with BERT-based embeddings. In 2022 *IEEE International Conference on Big Data* (*Big Data*) 5558–5567. IEEE.

SIBSON, R. (1969/70). Information radius. *Z. Wahrsch. Verw. Gebiete* **14** 149–160. MR0258198 https://doi.org/10.1007/BF00537520

SRIDHAR, D., DAUMÉ III, H. and BLEI, D. (2022). Heterogeneous supervised topic models. *Trans. Assoc. Comput. Linguist.* **10** 732–745.

SRIVASTAVA, A. and SUTTON, C. (2017). Autoencoding variational inference for topic models. Preprint. Available at arXiv:1703.01488.

THOMA, B. C., REZEPPA, T. L., CHOUKAS-BRADLEY, S., SALK, R. H. and MARSHAL, M. P. (2021). Disparities in childhood abuse between transgender and cisgender adolescents. *Pediatrics* **148**. https://doi.org/10.1542/peds.2020-016907

THOMPSON, L. and MIMNO, D. (2020). Topic modeling with contextualized word representation clusters. Preprint. Available at arXiv:2010.12626.

VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**.

WANG, C., BLEI, D. and HECKERMAN, D. (2012). Continuous time dynamic topic models. Preprint. Available at arXiv:1206.3298.

WANG, R., ZHOU, D. and HE, Y. (2019). Atm: Adversarial-neural topic model. *Inf. Process. Manag.* **56** 102098.

WHAIBEH, E., VOGT, E. L. and MAHMOUD, H. (2022). Addressing the behavioral health needs of sexual and gender minorities during the COVID-19 pandemic: A review of the expanding role of digital health technologies. *Curr. Psychiatry Rep.* **24** 387–397. https://doi.org/10.1007/s11920-022-01352-1

WU, T., JIA, X., SHI, H., NIU, J., YIN, X., XIE, J. and WANG, X. (2021). Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis. *J. Affective Disorders* **281** 91–98.

YE, H., MORENO, T., ALPERN, A., EHWERHEMUEPHA, L. and QU, A. (2024). Supplement to "Dynamic topic language model on heterogeneous children's mental health clinical notes." https://doi.org/10.1214/24-AOAS1930SUPP

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3185–3204
https://doi.org/10.1214/24-AOAS1931

RELIABILITY STUDY OF BATTERY LIVES: A FUNCTIONAL DEGRADATION ANALYSIS APPROACH

BY YOUNGJIN CHO[1,a], QUYEN DO[2,d], PANG DU[1,b] AND YILI HONG[1,c]

[1]*Department of Statistics, Virginia Tech,* [a]*youngjin@vt.edu,* [b]*pangdu@vt.edu,* [c]*yilihong@vt.edu*
[2]*Statistical Engineer, Corning Inc.,* [d]*DoQN@corning.com*

Renewable energy is critical for combating climate change, whose first step is the storage of electricity generated from renewable energy sources. Li-ion batteries are a popular kind of storage units. Their continuous usage through charge-discharge cycles eventually leads to degradation. This can be visualized by plotting voltage discharge curves (VDCs) over discharge cycles. Studies of battery degradation have mostly concentrated on modeling degradation through one scalar measurement summarizing each VDC. Such simplification of curves can lead to inaccurate predictive models. Here we analyze the degradation of rechargeable Li-ion batteries from a NASA data set through modeling and predicting their full VDCs. With techniques from longitudinal and functional data analysis, we propose a new two-step predictive modeling procedure for functional responses residing on heterogeneous domains. We first predict the shapes and domain end points of VDCs using functional regression models. Then we integrate these predictions to perform a degradation analysis. Our functional approach allows the incorporation of usage information, produces predictions in a curve form and thus provides flexibility in the assessment of battery degradation. Through extensive simulation studies and cross-validated data analysis, our approach demonstrates better prediction than the existing approach of modeling degradation directly with aggregated data.

REFERENCES

ANEIROS, G., CAO, R., FRAIMAN, R., GENEST, C. and VIEU, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *J. Multivariate Anal.* **170** 3–9. MR3913024 https://doi.org/10.1016/j.jmva.2018.11.007

BULL, S. R. (2001). Renewable energy today and tomorrow. *Proc. IEEE* **89** 1216–1226.

CARROLL, C., GAJARDO, A., 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 0.5.6.

CASTELVECCHI, D. (2021). Electric cars and batteries: How will the world produce enough? *Nature* **596** 336–339. https://doi.org/10.1038/d41586-021-02222-1

CHO, Y., DO, Q., DU, P. and HONG, Y. (2024). Supplement to "Reliability Study of Battery Lives: a Functional Degradation Analysis Approach." https://doi.org/10.1214/24-AOAS1931SUPP

DIOUF, B. and PODE, R. (2015). Potential of lithium-ion batteries in renewable energy. *Renew. Energy* **76** 375–380.

DUAN, Y., HONG, Y., MEEKER, W. Q., STANLEY, D. L. and GU, X. (2017). Photodegradation modeling based on laboratory accelerated test data and predictions under outdoor weathering for polymeric materials. *Ann. Appl. Stat.* **11** 2052–2079. MR3743288 https://doi.org/10.1214/17-AOAS1060

FANG, G. and PAN, R. (2024). A class of hierarchical multivariate Wiener processes for modeling dependent degradation data. *Technometrics* **66** 141–156. MR4740741 https://doi.org/10.1080/00401706.2023.2242413

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050 https://doi.org/10.1111/j.0006-341X.2002.00121.x

HE, Y.-J., SHEN, J.-N., SHEN, J.-F. and MA, Z.-F. (2015). State of health estimation of lithium-ion batteries: A multiscale Gaussian process regression modeling approach. *AIChE J.* **61** 1589–1600.

Key words and phrases. Degradation analysis, functional data analysis, longitudinal data, rechargeable batteries.

HONG, Y., DUAN, Y., MEEKER, W. Q., STANLEY, D. L. and GU, X. (2015). Statistical methods for degradation data with dynamic covariates information and an application to outdoor weathering data. *Technometrics* **57** 180–193. MR3369675 https://doi.org/10.1080/00401706.2014.915891

LI, Y., QIU, Y. and XU, Y. (2022). From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. *J. Multivariate Anal.* **188** Paper No. 104806, 15. MR4353837 https://doi.org/10.1016/j.jmva.2021.104806

LIU, B., WANG, L. and CAO, J. (2017). Estimating functional linear mixed-effects regression models. *Comput. Statist. Data Anal.* **106** 153–164. MR3566081 https://doi.org/10.1016/j.csda.2016.09.009

LIU, D., LUO, Y., LIU, J., PENG, Y., GUO, L. and PECHT, M. (2014). Lithium-ion battery remaining useful life estimation based on fusion nonlinear degradation AR model and RPF algorithm. *Neural Comput. Appl.* **25** 557–572.

LIU, D., LUO, Y., PENG, Y., PENG, X. and PECHT, M. (2012a). Lithium-ion battery remaining useful life estimation based on nonlinear AR model combined with degradation feature. In *Annual Conference of the PHM Society* **4**.

LIU, D., PANG, J., ZHOU, J. and PENG, Y. (2012b). Data-driven prognostics for lithium-ion battery based on Gaussian process regression. In *Proceedings of the IEEE* 2012 *Prognostics and System Health Management Conference* (*PHM*-2012 *Beijing*) 1–5. IEEE Press.

LIU, D., PANG, J., ZHOU, J., PENG, Y. and PECHT, M. (2013). Prognostics for state of health estimation of lithium-ion batteries based on combination Gaussian process functional regression. *Microelectron. Reliab.* **53** 832–839.

LIU, D., ZHOU, J., PAN, D., PENG, Y. and PENG, X. (2015). Lithium-ion battery remaining useful life estimation with an optimized relevance vector machine algorithm with incremental learning. *Measurement* **63** 143–151.

LUO, W., LV, C., WANG, L. and LIU, C. (2011). Study on impedance model of Li-ion battery. In 2011 6*th IEEE Conference on Industrial Electronics and Applications* 1943–1947. IEEE Press, New York.

MARTIN, J. A., OUWERKERK, J. N., LAMPING, A. P. and COHEN, K. (2022). Comparison of battery modeling regression methods for application to unmanned aerial vehicles. *Complex Eng. Syst.* **2**.

MEEKER, W., HONG, Y. and ESCOBAR, L. (2004). Degradation models and analyses. *Encycl. Stat. Sci.* 1–23.

MEEKER, W. Q., ESCOBAR, L. A. and PASCUAL, F. G. (2022). *Statistical Methods for Reliability Data*, 2nd ed. Wiley, Hoboken, NJ, USA.

NASCIMENTO, R. G., CORBETTA, M., KULKARNI, C. S. and VIANA, F. A. (2021). Hybrid physics-informed neural networks for lithium-ion battery modeling and prognosis. *J. Power Sources* **513** 230526.

NG, S. S., XING, Y. and TSUI, K. L. (2014). A naive Bayes model for robust remaining useful life prediction of lithium-ion battery. *Appl. Energy* **118** 114–123.

OLABI, A. and ABDELKAREEM, M. A. (2022). Renewable energy and climate change. *Renew. Sustain. Energy Rev.* **158** 112111.

PANWAR, N., KAUSHIK, S. and KOTHARI, S. (2011). Role of renewable energy sources in environmental protection: A review. *Renew. Sustain. Energy Rev.* **15** 1513–1524.

PATIL, M. A., TAGADE, P., HARIHARAN, K. S., KOLAKE, S. M., SONG, T., YEO, T. and DOO, S. (2015). A novel multistage support vector machine based approach for Li-ion battery remaining useful life estimation. *Appl. Energy* **159** 285–297.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D., HEISTERKAMP, S., VAN WILLIGEN, B. and MAINTAINER, R. (2017). R Package 'nlme' (version 3.1-135).

RAMSAY, J. O. and SILVERMAN, B. W. (1997). *Functional Data Analysis*, 1nd ed. Springer, New York. MR2168993

RICHARDSON, R. R., OSBORNE, M. A. and HOWEY, D. A. (2017). Gaussian process regression for forecasting battery state of health. *J. Power Sources* **357** 209–219.

SAHA, B. and GOEBEL, K. (2007). *Battery Data Set*. *NASA Prognostics Data Repository*. NASA Ames Research Center, Moffett Field, CA.

SAHA, B. and GOEBEL, K. (2009). Modeling Li-ion battery capacity depletion in a particle filtering framework. In *Annual Conference of the PHM Society* **1**.

SAHA, B., GOEBEL, K. and CHRISTOPHERSEN, J. (2009). Comparison of prognostic algorithms for estimating remaining useful life of batteries. *Trans. Inst. Meas. Control* **31** 293–308.

SAHA, B., GOEBEL, K., POLL, S. and CHRISTOPHERSEN, J. (2008). Prognostics methods for battery health monitoring using a Bayesian framework. *IEEE Trans. Instrum. Meas.* **58** 291–296.

SBARUFATTI, C., CORBETTA, M., GIGLIO, M. and CADINI, F. (2017). Adaptive prognosis of lithium-ion batteries based on the combination of particle filters and radial basis function neural networks. *J. Power Sources* **344** 128–140.

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

SHEN, Y., SHEN, L. and XU, W. (2018). A Wiener-based degradation model with logistic distributed measurement errors and remaining useful life estimation. *Qual. Reliab. Eng. Int.* **34** 1289–1303.

TAGADE, P., HARIHARAN, K. S., RAMACHANDRAN, S., KHANDELWAL, A., NAHA, A., KOLAKE, S. M. and HAN, S. H. (2020). Deep Gaussian process regression for lithium-ion battery health prognosis and degradation mode diagnosis. *J. Power Sources* **445** 227281.

TANG, S., YU, C., WANG, X., GUO, X. and SI, X. (2014). Remaining useful life prediction of lithium-ion batteries based on the Wiener process with measurement error. *Energies* **7** 520–547.

THIÉBAUT, R., JACQMIN-GADDA, H., CHÊNE, G., LEPORT, C. and COMMENGES, D. (2002). Bivariate linear mixed models using SAS proc MIXED. *Comput. Methods Programs Biomed.* **69** 249–256.

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

XU, L., GOTWALT, C., HONG, Y., KING, C. B. and MEEKER, W. Q. (2020). Applications of the fractional-random-weight bootstrap. *Amer. Statist.* **74** 345–358. MR4168255 https://doi.org/10.1080/00031305.2020.1731599

YE, Z.-S. and XIE, M. (2015). Stochastic modelling and analysis of degradation for highly reliable products. *Appl. Stoch. Models Bus. Ind.* **31** 16–32. MR3326372 https://doi.org/10.1002/asmb.2063

YU, Z., ZHANG, Y., QI, L. and LI, R. (2022). SOH estimation method for lithium-ion battery based on discharge characteristics. *Int. J. Electrochem. Sci.* **17**.

ZHOU, R. R., SERBAN, N. and GEBRAEEL, N. (2011). Degradation modeling applied to residual lifetime prediction using functional data analysis. *Ann. Appl. Stat.* **5** 1586–1610. MR2849787 https://doi.org/10.1214/10-AOAS448

ZHU, H., CHEN, K., LUO, X., YUAN, Y. and WANG, J.-L. (2019). FMEM: Functional mixed-effects models for longitudinal functional responses. *Statist. Sinica* **29** 2007–2033. MR3970345

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3205–3228
https://doi.org/10.1214/24-AOAS1933

LEARNING RISK PREFERENCES IN MARKOV DECISION PROCESSES: AN APPLICATION TO THE FOURTH DOWN DECISION IN THE NATIONAL FOOTBALL LEAGUE

BY NATHAN SANDHOLTZ[1,a], LUCAS WU[2,b], MARTIN PUTERMAN[3,c] AND TIMOTHY C. Y. CHAN[4,d]

[1]*Department of Statistics, Brigham Young University,* [a]*nsandholtz@stat.byu.edu*
[2]*Zelus Analytics,* [b]*lwu@zelusanalytics.com*
[3]*Sauder School of Business, University of British Columbia,* [c]*martin.puterman@sauder.ubc.ca*
[4]*Department of Mechanical & Industrial Engineering, University of Toronto,* [d]*tcy.chan@utoronto.ca*

For decades National Football League (NFL) coaches' observed fourth down decisions have been largely inconsistent with prescriptions based on statistical models. In this paper we develop a framework to explain this discrepancy using an inverse optimization approach. We model the fourth down decision and the subsequent sequence of plays in a game as a Markov decision process (MDP), the dynamics of which we estimate from NFL play-by-play data from the 2014 through 2022 seasons. We assume that coaches' observed decisions are optimal but that the risk preferences governing their decisions are unknown. This yields an inverse decision problem for which the optimality criterion, or risk measure, of the MDP is the estimand. Using the quantile function to parameterize risk, we estimate which quantile-optimal policy yields the coaches' observed decisions as minimally suboptimal. In general, we find that coaches' fourth-down behavior is consistent with optimizing low quantiles of the next-state value distribution, which corresponds to conservative risk preferences. We also find that coaches exhibit higher risk tolerances when making decisions in the opponent's half of the field, as opposed to their own half, and that league average fourth down risk tolerances have increased over time.

REFERENCES

ALAMAR, B. C. (2006). The passing premium puzzle. *J. Quant. Anal. Sports* **2** Art. 5, 10. MR2270284 https://doi.org/10.2202/1559-0410.1051

ALAMAR, B. C. (2010). Measuring risk in NFL playcalling. *J. Quant. Anal. Sports* **6**.

BALDWIN, B. (2021a). 4th down research. https://www.nfl4th.com/articles/articles/4th-down-research.html. Accessed: 2021-03-31.

BALDWIN, B. (2021b). Open source football: NflfastR EP, WP, CP xYAC, and xPass models. https://www.opensourcefootball.com/posts/2020-09-28-nflfastr-ep-wp-and-cp-models/. Accessed: 2024-04-30.

BALDWIN, B. (2024). nfl4th: Functions to calculate optimal fourth down decisions in the national football league.

BELLEMARE, M. G., DABNEY, W. and MUNOS, R. (2017). A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* 449–458. PMLR.

BRILL, R. S., YURKO, R. and WYNER, A. J. (2023). Analytics, have some humility: A statistical view of fourth-down decision making. Preprint. Available at arXiv:2311.03490.

BURKE, B., CARTER, S., GIRATIKANON, T., QUEALY, K. and DANIEL, J. (2014). 4th down: When to go for it and why. https://www.nytimes.com/2014/09/05/upshot/4th-down-when-to-go-for-it-and-why.html. Accessed: 2021-03-23.

CARL, S. and BALDWIN, B. (2024). nflfastR: Functions to efficiently access NFL play by play data.

CARTER, V. and MACHOL, R. E. (1971). Operations research on football. *Oper. Res.* **19** 541–544.

CARTER, V. and MACHOL, R. E. (1978). Note—optimal strategies on fourth down. *Manage. Sci.* **24** 1758–1762.

CHAN, T. C., MAHMOOD, R. and ZHU, I. Y. (2023). Inverse optimization: Theory and applications. *Oper. Res.* **0**.

CHAN, T. C. Y., FERNANDES, C. and PUTERMAN, M. L. (2021). Points gained in football: Using Markov process-based value functions to assess team performance. *Oper. Res.* **69** 877–894. MR4280422 https://doi.org/10.1287/opre.2020.2034

Key words and phrases. Inverse optimization, MDP, quantile function, decision analysis, sport.

CRITCHFIELD, T. S. and STILLING, S. T. (2015). A matching law analysis of risk tolerance and gain–loss framing in football play selection. *Behavior Analysis*: *Research and Practice* **15** 112.
DALY-GRAFSTEIN, D. (2023). Correcting for preferential bias in NFL fourth-down decision making. In *Proceedings of the* 17*th MIT Sloan Sports Analytics Conference*, *Boston*, *MA*, *USA*.
DI CASTRO, D., OREN, J. and MANNOR, S. (2019). Practical risk measures in reinforcement learning. Preprint. Available at arXiv:1908.08379.
EFRON, B. (1982). *The Jackknife*, *the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. SIAM, Philadelphia, PA. MR0659849
VON NEUMANN, J. and MORGENSTERN, O. (2007). Theory of games and economic behavior. In *Theory of Games and Economic Behavior* Princeton university press.
FRANKS, A. M., D'AMOUR, A., CERVONE, D. and BORNN, L. (2016). Meta-analytics: Tools for understanding the statistical properties of sports metrics. *J. Quant. Anal. Sports* **12** 151–165.
GILBERT, H., WENG, P. and XU, Y. (2017). Optimizing quantiles in preference-based Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence* **31**.
GOLDNER, K. (2012). A Markov model of football: Using stochastic processes to model a football drive. *J. Quant. Anal. Sports* **8**.
HAMMING, R. W. (1950). Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29** 147–160. MR0035935 https://doi.org/10.1002/j.1538-7305.1950.tb00463.x
JAQUETTE, S. C. (1973). Markov decision processes with a new optimality criterion: Discrete time. *Ann. Statist.* **1** 496–505. MR0378839
KAHNEMAN, D. and TVERSKY, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making*: *Part I* 99–127. World Scientific, Singapore.
KOVASH, K. and LEVITT, S. D. (2009). Professionals do not play minimax: Evidence from major league baseball and the national football league Technical Report National Bureau of Economic Research.
LI, J. Y.-M. (2021). Inverse optimization of convex risk functions. *Manage. Sci.* **67** 7113–7141.
LI, X., ZHONG, H. and BRANDEAU, M. L. (2022). Quantile Markov decision processes. *Oper. Res.* **70** 1428–1447. MR4451046 https://doi.org/10.1287/opre.2021.2123
LOPEZ, M. J. (2020). Bigger data, better questions, and a return to fourth down behavior: An introduction to a special issue on tracking datain the national football league. *J. Quant. Anal. Sports* **16** 73–79.
MASSEY, C. and THALER, R. H. (2013). The loser's curse: Decision making and market efficiency in the national football league draft. *Manage. Sci.* **59** 1479–1495.
OWENS, M. F. and ROACH, M. A. (2018). Decision-making on the hot seat and the short list: Evidence from college football fourth down decisions. *J. Econ. Behav. Organ.* **148** 301–314.
PYA, N. (2020). scam: Shape constrained additive models. R package version 1.2-9.
PYA, N. and WOOD, S. N. (2015). Shape constrained additive models. *Stat. Comput.* **25** 543–559. MR3334416 https://doi.org/10.1007/s11222-013-9448-7
ROACH, M. A. and OWENS, M. F. (2024). Updating beliefs based on observed performance: Evidence from NFL head coaches. *J. Sports Econ.* **25** 369–387.
ROMER, D. (2006). Do firms maximize? Evidence from professional football. *Eur. J. Polit. Econ.* **114** 340–365.
SANDHOLTZ, N., MIYAMOTO, Y., BORNN, L. and SMITH, M. A. (2023). Inverse Bayesian optimization: Learning human acquisition functions in an exploration vs exploitation search task. *Bayesian Anal.* **18** 1–24. MR4515723 https://doi.org/10.1214/21-ba1303
SANDHOLTZ, N., WU, L., PUTERMAN, M. and CHAN, T. C (2024). Supplement to "Learning risk preferences in Markov decision processes: An application to the fourth down decision in the national football league." https://doi.org/10.1214/24-AOAS1933SUPPA, https://doi.org/10.1214/24-AOAS1933SUPPB
NG, A. Y., RUSSELL, S. J. et al. (2000). Algorithms for inverse reinforcement learning. In *Icml* **1** 2.
SOBEL, M. J. (1982). The variance of discounted Markov decision processes. *J. Appl. Probab.* **19** 794–802. MR0675143 https://doi.org/10.2307/3213832
URSCHEL, J. D. and ZHUANG, J. (2011). Are NFL coaches risk and loss averse? Evidence from their use of kickoff strategies. *J. Quant. Anal. Sports* **7**.
VARIAN, H. R. (2006). Revealed preference. *Samuelsonian Economics and the Twenty-first Century* 99–115.
WHITE, D. J. (1988). Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *J. Optim. Theory Appl.* **56** 1–29. MR0922375 https://doi.org/10.1007/BF00938524
WINSTON, W., CABOT, V. and SAGARIN, J. (1983). Football as an infinite horizon zero-sum stochastic game. Technical Report.
YAM, D. R. and LOPEZ, M. J. (2019). What was lost? A causal estimate of fourth down behavior in the national football league. *J. Sports Anal.* **5** 153–167.
YU, S., CHEN, Y. and DONG, C. (2020). Learning time varying risk preferences from investment portfolios using inverse optimization with applications on mutual funds. Preprint. Available at arXiv:2010.01687.

YURKO, R., VENTURA, S. and HOROWITZ, M. (2019). nflwar: A reproducible method for offensive player evaluation in football. *J. Quant. Anal. Sports* **15** 163–183.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3229–3252
https://doi.org/10.1214/24-AOAS1934

EXTENDED BETA MODELS FOR POVERTY MAPPING. AN APPLICATION INTEGRATING SURVEY AND REMOTE SENSING DATA IN BANGLADESH

BY SILVIA DE NICOLÒ[1,a], ENRICO FABRIZI[2] AND ALDO GARDINI[1]

[1]*Department of Statistical Sciences, Università di Bologna,* [a]*silvia.denicolo@unibo.it*
[2]*DISES, Università Cattolica del S. Cuore*

The paper targets the estimation of a poverty rate at the upazila level in Bangladesh through the use of demographic and health survey (DHS) data. Upazilas are administrative regions equivalent to counties or boroughs whose sample sizes are not large enough to provide reliable estimates or are even absent. We tackle this issue by proposing a small-area estimation model complementing survey data with remote sensing information at the area level. We specify an extended Beta mixed regression model within the Bayesian framework, allowing it to accommodate the peculiarities of sample data and to predict out-of-sample rates. Specifically, it enables to include estimates equal to either 0 or 1 and to model the strong intra-cluster correlation. We aim at proposing a method that can be implemented by statistical offices as a routine. In this spirit we consider a regularizing prior for coefficients, rather than a model selection approach, to deal with a large number of auxiliary variables. We compare our methods with existing alternatives using a design-based simulation exercise and illustrate its potential with the motivating application.

REFERENCES

BENEDETTI, M. H., BERROCAL, V. J. and LITTLE, R. J. (2022). Accounting for survey design in Bayesian disaggregation of survey-based areal estimates of proportions: An application to the American Community Survey. *Ann. Appl. Stat.* **16** 2201–2230. MR4489206 https://doi.org/10.1214/21-aoas1585

BOUBETA, M., LOMBARDÍA, M. J. and MORALES, D. (2017). Poisson mixed models for studying the poverty in small areas. *Comput. Statist. Data Anal.* **107** 32–47. MR3575057 https://doi.org/10.1016/j.csda.2016.10.014

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

BROWN, P. J. and GRIFFIN, J. E. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440 https://doi.org/10.1214/10-BA507

BURGERT, C. R., ZACHARY, B. and COLSTON, J. (2013). *Incorporating Geographic Information into Demographic and Health Surveys*: *A Field Guide to GPs Data Collection*. ICF International, Calverton, MD, USA.

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**.

CASAS-CORDERO VALENCIA, C., ENCINA, J. and LAHIRI, P. (2016). Poverty mapping for the Chilean comunas. In *Analysis of Poverty Data by Small Area Estimation* 379–404. Wiley, New York.

CHEN, C., WAKEFIELD, J. and LUMELY, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat. Spatio-Tempor. Epidemiol.* **11** 33–43. https://doi.org/10.1016/j.sste.2014.07.002

CHEN, S. and RUST, K. (2017). An extension of Kish's formula for design effects to two-and three-stage designs with stratification. *J. Surv. Stat. Methodol.* **5** 111–130.

CORSI, D. J., NEUMAN, M., FINLAY, J. E. and SUBRAMANIAN, S. V. (2012). Demographic and health surveys: A profile. *Int. J. Epidemiol.* **41** 1602–1613. https://doi.org/10.1093/ije/dys184

DATTA, G. S., HALL, P. and MANDAL, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *J. Amer. Statist. Assoc.* **106** 362–374. MR2816727 https://doi.org/10.1198/jasa.2011.tm10036

Key words and phrases. Demographic health survey, hierarchical Bayes, shrinkage priors, small area estimation.

DE NICOLÒ, S., FABRIZI, E. and GARDINI, A. (2024). Supplement to "Extended Beta models for poverty mapping. An application integrating survey and remote sensing data in Bangladesh." https://doi.org/10.1214/24-AOAS1934SUPPA, https://doi.org/10.1214/24-AOAS1934SUPPB

DE NICOLÒ, S. and GARDINI, A. (2024). The R package tipsae: Tools for mapping proportions and indicators on the unit interval. *J. Stat. Softw.* **108** 1–36.

DURANTON, G. and VENABLES, A. J. (2021). Place-based policies: Principles and developing country applications. In *Handbook of Regional Science* 1009–1030. Springer, Berlin.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319. MR0391403

ENGSTROM, R., HERSH, J. S. and NEWHOUSE, D. L. (2017). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. World Bank Policy Research Working Paper 8284.

FABRIZI, E., FERRANTE, M. and TRIVISANO, C. (2016). Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. In *Analysis of Poverty Data by Small Area Methods* 299–314. Wiley, New York.

FABRIZI, E., FERRANTE, M. R., PACEI, S. and TRIVISANO, C. (2011). Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Comput. Statist. Data Anal.* **55** 1736–1747. MR2748675 https://doi.org/10.1016/j.csda.2010.11.001

FABRIZI, E., FERRANTE, M. R. and TRIVISANO, C. (2018). Bayesian small area estimation for skewed business survey variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **67** 861–879. MR3832254 https://doi.org/10.1111/rssc.12254

FABRIZI, E., FERRANTE, M. R. and TRIVISANO, C. (2020). A functional approach to small area estimation of the relative median poverty gap. *J. Roy. Statist. Soc. Ser. A* **183** 1273–1291. MR4114486 https://doi.org/10.1111/rssa.12562

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

FERRARI, S. L. P. and CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31** 799–815. MR2095753 https://doi.org/10.1080/0266476042000214501

FRANCO, C. and BELL, W. R. (2015). Borrowing Information over Time in Binomial/Logit Normal Models for Small Area Estimation. *Stat. Transit.* **4** 563–584.

FULLER, W. A. (2011). *Sampling Statistics*. Wiley, New York.

GABLER, S., HÄDER, S. and LAHIRI, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Surv. Methodol.* **25** 105–106.

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

HÁJEK, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. Foundations of Statistical Inference 326. MR0423625

HAQUE, A. and JAHAN, S. (2015). Impact of flood disasters in Bangladesh: A multi-sector regional analysis. *Int. J. Disaster Risk Reduct.* **13** 266–275.

HAY, S. I. and SNOW, R. W. (2006). The malaria Atlas Project: Developing global maps of malaria risk. *PLoS Med.* **3** e473. https://doi.org/10.1371/journal.pmed.0030473

IIMI, A., AHMED, F., ANDERSON, E. C., DIEHL, A. S., MAIYO, L., PERALTA-QUIRÓS, T. and RAO, K. (2016). New rural access index: Main determinants and correlation to poverty. World Bank Policy Research Working Paper 7876.

IMAM, M. F., ISLAM, M. A., ALAM, M. A., HOSSAIN, M. J. and DAS, S. (2019). Small area estimation of poverty in rural Bangladesh. *Bangladesh J. Agric. Econ.* **40** 1–16.

ISLAM, D., SAYEED, J. and HOSSAIN, N. (2017). On determinants of poverty and inequality in Bangladesh. *J. Poverty* **21** 352–371.

JANICKI, R. (2020). Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Comm. Statist. Theory Methods* **49** 2264–2284. MR4075493 https://doi.org/10.1080/03610926.2019.1570266

KALTON, G. (1979). Ultimate cluster sampling. *J. Roy. Statist. Soc. Ser. A* **142** 210–222. MR0547238 https://doi.org/10.2307/2345081

KAM, S.-P., HOSSAIN, M., BOSE, M. L. and VILLANO, L. S. (2005). Spatial patterns of rural poverty and their relationship with welfare-influencing factors in Bangladesh. *Food Policy* **30** 551–567.

KHUDRI, M. M., CHOWDHURY, F. et al. (2013). Evaluation of socio-economic status of households and identifying key determinants of poverty in Bangladesh. *Eur. J. Soc. Sci.* **37** 377–387.

KISH, L. (1987). Weighting in Deft2. *Surv. Stat.* **17** 26–30.

KLOTZ, J. (1973). Statistical inference in Bernoulli trials with dependence. *Ann. Statist.* **1** 373–379. MR0381103

KREUTZMANN, A.-K., PANNIER, S., ROJAS-PERILLA, N., SCHMID, T., TEMPL, M. and TZAVIDIS, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *J. Stat. Softw.* **91** 1–33. https://doi.org/10.18637/jss.v091.i07

LIU, B., LAHIRI, P. and KALTON, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. In *Proceedings of the American Statistical Association*, *Survey Research Section* 3181–3186.

LYNN, P., HÄDER, S. and GABLER, S. (2006). Design effects for multiple design samples. *Surv. Methodol.* **32** 115–120.

MARHUENDA, Y., MOLINA, I. and MORALES, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Comput. Statist. Data Anal.* **58** 308–325. MR2997945 https://doi.org/10.1016/j.csda.2012.09.002

MASAKI, T., NEWHOUSE, D., SILWAL, A. R., BEDADA, A. and ENGSTROM, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Stat. J. IAOS* **38** 1035–1051.

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

NIPORT AND MITRA AND ASSOCIATES AND ICF INTERNATIONAL (2016). Bangladesh Demographic and Health Survey 2014. National Institute of Population Research and Training (NIPORT), Mitra and Associates, and ICF International, Dhaka, Bangladesh, and Rockville, Maryland, USA.

O'DONNELL, M. S. and IGNIZIO, D. A. (2012). Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States. US Geological Survey Data Series No. 691.

PEZZULO, C., HORNBY, G. M., SORICHETTA, A., GAUGHAN, A. E., LINARD, C., BIRD, T. J., KERR, D., LLOYD, C. T. and TATEM, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. *Sci. Data* **4** 1–15.

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

POIRIER, M. J., GRÉPIN, K. A. and GRIGNON, M. (2020). Approaches and alternatives to the wealth index to measure socioeconomic status using survey data: A critical interpretive synthesis. *Soc. Indic. Res.* **148** 1–46.

PORTER, A. T., HOLAN, S. H., WIKLE, C. K. and CRESSIE, N. (2014). Spatial Fay-Herriot models for small area estimation with functional covariates. *Spat. Stat.* **10** 27–42. MR3280088 https://doi.org/10.1016/j.spasta.2014.07.001

RAGHUNATHAN, T. E., XIE, D., SCHENKER, N., PARSONS, V. L., DAVIS, W. W., DODD, K. W. and FEUER, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *J. Amer. Statist. Assoc.* **102** 474–486. MR2370848 https://doi.org/10.1198/016214506000001293

RAHMAN, M. (2017). Role of agriculture in Bangladesh economy: Uncovering the problems and challenges. *Int. J. Bus. Manag. Invent.* **6**.

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

RIDOUT, M. S., DEMÉTRIO, C. G. and FIRTH, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55** 137–148. https://doi.org/10.1111/j.0006-341x.1999.00137.x

SCHMID, T., BRUCKSCHEN, F., SALVATI, N. and ZBIRANSKI, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *J. Roy. Statist. Soc. Ser. A* **180** 1163–1190. MR3723791 https://doi.org/10.1111/rssa.12305

STEELE, J. E., SUNDSØY, P. R., PEZZULO, C., ALEGANA, V. A., BIRD, T. J., BLUMENSTOCK, J., BJELLAND, J., ENGØ-MONSEN, K., DE MONTJOYE, Y.-A. et al. (2017). Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* **14** 20160690.

SUGASAWA, S. and KUBOKAWA, T. (2017). Transforming response values in small area prediction. *Comput. Statist. Data Anal.* **114** 47–60. MR3660838 https://doi.org/10.1016/j.csda.2017.03.017

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

TATEM, A. J. (2017). WorldPop, open data for spatial demography. *Sci. Data* **4** 1–4.

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

WARTON, D. I. and HUI, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology* **92** 3–10. https://doi.org/10.1890/10-0340.1

WIECZOREK, J. and HAWALA, S. (2011). A Bayesian zero-one inflated beta model for estimating poverty in US counties. In *Proceedings of the American Statistical Association*, *Section on Survey Research Methods* Amer. Statist. Assoc., Alexandria, VA.

WORLD BANK (2008). Poverty Assessment for Bangladesh: Creating Opportunities and Bridging the East-West Divide. Bangladesh Development Series Paper No. 26.

XIE, D., RAGHUNATHAN, T. E. and LEPKOWSKI, J. M. (2007). Estimation of the proportion of overweight individuals in small areas—a robust extension of the Fay-Herriot model. *Stat. Med.* **26** 2699–2715. MR2370833 https://doi.org/10.1002/sim.2709

ZHAO, X., YU, B., LIU, Y., CHEN, Z., LI, Q., WANG, C. and WU, J. (2019). Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. *Remote Sens.* **11** 375.

ZHOU, Y., MA, T., ZHOU, C. and XU, T. (2015). Nighttime light derived assessment of regional inequality of socioeconomic development in China. *Remote Sens.* **7** 1242–1262.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3253–3273
https://doi.org/10.1214/24-AOAS1935

A LATENT VARIABLE MIXTURE MODEL FOR COMPOSITION-ON-COMPOSITION REGRESSION WITH APPLICATION TO CHEMICAL RECYCLING

BY NICHOLAS RIOS[1,a], LINGZHOU XUE[2,b] AND XIANG ZHAN[3,c]

[1]*Department of Statistics, George Mason University,* [a]*nrios4@gmu.edu*
[2]*Department of Statistics, Pennsylvania State University,* [b]*lzxue@psu.edu*
[3]*Department of Biostatistics and Beijing International Center for Mathematical Research, Peking University,* [c]*zhanx@bjmu.edu.cn*

It is quite common to encounter compositional data in a regression framework in data analysis. When both responses and predictors are compositional, most existing models rely on a family of log-ratio based transformations to move the analysis from the simplex to the reals. This often makes the interpretation of the model more complex. A transformation-free regression model was recently developed, but it only allows for a single compositional predictor. However, many datasets include multiple compositional predictors of interest. Motivated by an application to hydrothermal liquefaction (HTL) data, a novel extension of this transformation-free regression model is provided that allows for two (or more) compositional predictors to be used via a latent variable mixture. A modified expectation-maximization algorithm is proposed to estimate model parameters, which are shown to have natural interpretations. Conformal inference is used to obtain prediction limits on the compositional response. The resulting methodology is applied to the HTL dataset. Extensions to multiple predictors are discussed.

REFERENCES

AGARWAL, A. and XUE, L. (2020). Model-based clustering of nonparametric weighted networks with application to water pollution analysis. *Technometrics* **62** 161–172. MR4095742 https://doi.org/10.1080/00401706.2019.1623076

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

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

RIOS, N., XUE, L. and ZHAN, X. (2024). Supplement to "A Latent Variable Mixture Model for Composition-on-Composition Regression with Application to Chemical Recycling." https://doi.org/10.1214/24-AOAS1935SUPPA, https://doi.org/10.1214/24-AOAS1935SUPPB

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, J., ZHANG, X. and LI, S. (2017). Multiple linear regression with compositional response and covariates. *J. Appl. Stat.* **44** 2270–2285. MR3670303 https://doi.org/10.1080/02664763.2016.1157145

CHERNOZHUKOV, V., WÜTHRICH, K. and YINCHU, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference on Learning Theory* 732–749. PMLR.

DESARBO, W. S., CHEN, Q. and BLANK, A. S. (2017). A parametric constrained segmentation methodology for application in sport marketing. *Cust. Needs Solut.* **4** 37–55.

DOUMA, J. C. and WEEDON, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol. Evol.* **10** 1412–1430.

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. and BARCELÓ-VIDAL, C. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* **35** 279–300. MR1986165 https://doi.org/10.1023/A:1023818214614

FIKSEL, J., ZEGER, S. and DATTA, A. (2022). A transformation-free linear regression for compositional outcomes and predictors. *Biometrics* **78** 974–987. MR4493502 https://doi.org/10.1111/biom.13465

Key words and phrases. Compositional data, EM-algorithm, Kullback–Leibler distance, transformation-free.

GOLLAKOTA, A., KISHORE, N. and GU, S. (2018). A review on hydrothermal liquefaction of biomass. *Renew. Sustain. Energy Rev.* **81** 1378–1392.

GUIRGUIS, P. M., SESHASAYEE, M. S., MOTAVAF, B. and SAVAGE, P. E. (2024). Review and assessment of models for predicting biocrude yields from hydrothermal liquefaction of biomass. *RSC Sustain.* **2** 736–756.

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

LEE, K. H., CHEN, Q., DESARBO, W. S. and XUE, L. (2022). Estimating finite mixtures of ordinal graphical models. *Psychometrika* **87** 83–106. MR4410418 https://doi.org/10.1007/s11336-021-09781-2

LEE, K. H. and XUE, L. (2018). Nonparametric finite mixture of Gaussian graphical models. *Technometrics* **60** 511–521. MR3878105 https://doi.org/10.1080/00401706.2017.1408497

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

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917 https://doi.org/10.1093/biomet/asu031

LU, J., LIU, Z., ZHANG, Y. and SAVAGE, P. E. (2018). Synergistic and antagonistic interactions during hydrothermal liquefaction of soybean oil, soy protein, cellulose, xylose, and lignin. *ACS Sustain. Chem. Eng.* **6** 14501–14509.

MAHADEVAN, S., RIOS, N., KOLLAR, A. J., STOFANAK, R., MALONEY, K., WALTZ, K. E., RANE, C., ENDLURI, S. and SAVAGE, P. E. (2023). Dataset for oil composition, yield from Hydrothermal Liquefaction of biomass. Mendeley Data 1. https://doi.org/10.17632/s38wv3fvpz.1

MCKENDRY, P. (2002). Energy production from biomass (part 1): Overview of biomass. *Bioresour. Technol.* **83** 37–46. https://doi.org/10.1016/s0960-8524(01)00118-3

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

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

PAWLOWSKY-GLAHN, V. and EGOZCUE, J. J. (2006). *Compositional Data and Their Analysis*: *An Introduction*. *Special Publications* **264** 1–10. Geological Society, London.

PEARSON, K. (1897). On a form of spurious correlation which may arise when indices are useed in the measurement of organs. *Proc. R. Soc. Lond.* **60** 489–502.

SHAFER, G. and VOVK, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9** 371–421. MR2417240

SHAHBEIK, H., PANAHI, H. K. S., DEHHAGHI, M., GUILLEMIN, G. J., FALLAHI, A., HOSSEINZADEH-BANDBAFHA, H., AMIRI, H., REHAN, M., RAIKWAR, D. et al. (2024). Biomass to biofuels using hydrothermal liquefaction: A comprehensive review. *Renew. Sustain. Energy Rev.* **189** 113976.

SRINIVASAN, A., XUE, L. and ZHAN, X. (2021). Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics* **77** 984–995. MR4320672 https://doi.org/10.1111/biom.13336

SRINIVASAN, A., XUE, L. and ZHAN, X. (2023). Identification of microbial features in multivariate regression under false discovery rate control. *Comput. Statist. Data Anal.* **181** Paper No. 107621, 11. MR4545143 https://doi.org/10.1016/j.csda.2022.107621

SUBRAMANYA, S. M., RIOS, N., KOLLAR, A., STOFANAK, R., MALONEY, K., WALTZ, K., POWERS, L., RANE, C. and SAVAGE, P. E. (2023). Statistical models for predicting oil composition from hydrothermal liquefaction of biomass. *Energy Fuels* **37** 6619–6628.

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

VALDEZ, P. J., TOCCO, V. J. and SAVAGE, P. E. (2014). A general kinetic model for the hydrothermal liquefaction of microalgae. *Bioresour. Technol.* **163** 123–127. https://doi.org/10.1016/j.biortech.2014.04.013

YU, X., YAO, J. and XUE, L. (2022). Nonparametric estimation and conformal inference of the sufficient forecasting with a diverging number of factors. *J. Bus. Econom. Statist.* **40** 342–354. MR4356577 https://doi.org/10.1080/07350015.2020.1813589

ZHANG, L., DOU, X., YANG, Z., YANG, X. and GUO, X. (2021). Advance in hydrothermal bio-oil preparation from lignocellulose: Effect of raw materials and their tissue structures. *Biomass* **1** 74–93.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3274–3296
https://doi.org/10.1214/24-AOAS1936

BAYESIAN ROBUST LEARNING IN CHAIN GRAPH MODELS FOR INTEGRATIVE PHARMACOGENOMICS

BY MOUMITA CHAKRABORTY[1,a], VEERABHADRAN BALADANDAYUTHAPANI[2,b], ANINDYA BHADRA[3,c] AND MIN JIN HA[4,d]

[1]*Department of Biostatistics and Data Science, The University of Texas Medical Branch,* [a]*mochakra@utmb.edu*
[2]*Department of Biostatistics, University of Michigan,* [b]*veerab@umich.edu*
[3]*Department of Statistics, Purdue University,* [c]*bhadra@purdue.edu*
[4]*Department of Health Informatics and Biostatistics, Yonsei University,* [d]*mjha@yuhs.ac*

Integrative analysis of multilevel pharmacogenomic data for modeling dependencies across various biological domains is crucial for developing genomic-testing based treatments. Chain graphs characterize conditional dependence structures of such multilevel data where variables are naturally partitioned into multiple ordered layers, consisting of both directed and undirected edges. Existing literature mostly focus on Gaussian chain graphs, which are ill-suited for nonnormal distributions with heavy-tailed marginals, potentially leading to inaccurate inferences. We propose a Bayesian robust chain graph model (RCGM) based on random transformations of marginals using Gaussian scale mixtures to account for node-level nonnormality in continuous multivariate data. This flexible modeling strategy facilitates identification of conditional sign dependencies among nonnormal nodes while still being able to infer conditional dependencies among normal nodes. In simulations we demonstrate that RCGM outperforms existing Gaussian chain graph inference methods in data generated from various nonnormal mechanisms. We apply our method to genomic, transcriptomic and proteomic data to understand underlying biological processes holistically for drug response and resistance in lung cancer cell lines. Our analysis reveals inter- and intra-platform dependencies of key signaling pathways to monotherapies of icotinib, erlotinib and osimertinib among other drugs, along with shared patterns of molecular mechanisms behind drug actions.

REFERENCES

AKBANI, R., NG, P. K. S., WERNER, H. M., SHAHMORADGOLI, M., ZHANG, F., JU, Z., LIU, W., YANG, J.-Y., YOSHIHARA, K. et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat. Commun.* **5** 1–15.

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. MR1844349 https://doi.org/10.1111/1467-9469.00224

ARMSTRONG, H. J. (2005). Bayesian estimation of decomposable Gaussian graphical models, PhD thesis, Univ. New South Wales.

BALDI, A., LUCA, A. D., ESPOSITO, V., CAMPIONI, M., SPUGNINI, E. P. and CITRO, G. (2011). Tumor suppressors and cell-cycle proteins in lung cancer. *For. Pathol.* **2011** 605042. https://doi.org/10.4061/2011/605042

BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHÁR, J., KRYUKOV, G. V. et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483** 603–607.

BEDARD, P. L., HANSEN, A. R., RATAIN, M. J. and SIU, L. L. (2013). Tumour heterogeneity in the clinic. *Nature* **501** 355–364. https://doi.org/10.1038/nature12627

BHADRA, A. and MALLICK, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics* **69** 447–457. MR3071063 https://doi.org/10.1111/biom.12021

BHADRA, A., RAO, A. and BALADANDAYUTHAPANI, V. (2018). Inferring network structure in non-normal and mixed discrete-continuous genomic data. *Biometrics* **74** 185–195. MR3777939 https://doi.org/10.1111/biom.12711

Key words and phrases. Bayesian graphical models, cancer, data integration, robust graphical models, multi-platform genomics, pharmacogenomics.

BRAMBILLA, E. and GAZDAR, A. (2009). Pathogenesis of lung cancer signalling pathways: Roadmap for therapies. *Eur. Respir. J.* **33** 1485–1497. https://doi.org/10.1183/09031936.00014009

BURGESS, J. T., ROSE, M., BOUCHER, D., PLOWMAN, J., MOLLOY, C., FISHER, M., O'LEARY, C., RICHARD, D. J., O'BYRNE, K. J. et al. (2020). The therapeutic potential of DNA damage repair pathways and genomic stability in lung cancer. *Front. Oncol.* **10** 1256. https://doi.org/10.3389/fonc.2020.01256

CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935 https://doi.org/10.1214/18-AOS1689

CASTELLETTI, F. and MASCARO, A. (2021). Structural learning and estimation of joint causal effects among network-dependent variables. *Stat. Methods Appl.* **30** 1289–1314. MR4343024 https://doi.org/10.1007/s10260-021-00579-1

CASTELLETTI, F. and MASCARO, A. (2022). Bcdag: an R package for Bayesian structure and causal learning of Gaussian dags. arXiv preprint. Available at arXiv:2201.12003.

CHAKRABORTY, M., BALADANDAYUTHAPANI, V., BHADRA, A. and HA, M. J. (2024). Supplement to "Bayesian Robust Learning in Chain Graph Models for Integrative Pharmacogenomics." https://doi.org/10.1214/24-AOAS1936SUPP

CHEN, M., REN, Z., ZHAO, H. and ZHOU, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111** 394–406. MR3494667 https://doi.org/10.1080/01621459.2015.1010039

CONSONNI, G., LA ROCCA, L. and PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.* **44** 741–764. MR3687971 https://doi.org/10.1111/sjos.12273

CONTE, N., MASON, J. C., HALMAGYI, C., NEUHAUSER, S., MOSAKU, A., YORDANOVA, G., CHATZIPLI, A., BEGLEY, D. A., KRUPKE, D. M. et al. (2019). PDX finder: A portal for patient-derived tumor xenograft model discovery. *Nucleic Acids Res.* **47** D1073–D1079.

CORSELLO, S. M., NAGARI, R. T., SPANGLER, R. D., ROSSEN, J., KOCAK, M., BRYAN, J. G., HUMEIDI, R., PECK, D., WU, X. et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **1** 235–248.

DAVIES, M., HENNESSY, B. and MILLS, G. B. (2006). Point mutations of protein kinases and individualised cancer therapy. *Expert Opin. Pharmacother.* **7** 2243–2261. https://doi.org/10.1517/14656566.7.16.2243

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 https://doi.org/10.1214/aos/1176349260

DOBRA, A., LENKOSKI, A. et al. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183 https://doi.org/10.1214/10-AOAS397

DRTON, M. and EICHLER, M. (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Stat.* **33** 247–257. MR2279641 https://doi.org/10.1111/j.1467-9469.2006.00482.x

DRTON, M. and PERLMAN, M. D. (2008). A SINful approach to Gaussian graphical model selection. *J. Statist. Plann. Inference* **138** 1179–1200. MR2416875 https://doi.org/10.1016/j.jspi.2007.05.035

FINEGOLD, M. and DRTON, M. (2011). Robust graphical modeling of gene networks using classical and alternative t-distributions. *Ann. Appl. Stat.* **5** 1057–1080. MR2840186 https://doi.org/10.1214/10-AOAS410

FINEGOLD, M. and DRTON, M. (2014). Robust Bayesian graphical modeling using Dirichlet t-distributions. *Bayesian Anal.* **9** 521–550. MR3256052 https://doi.org/10.1214/13-BA856

FINK, L. S., LERNER, C. A., TORRES, P. F. and SELL, C. (2010). Ku80 facilitates chromatin binding of the telomere binding protein, TRF2. *Cell Cycle* **9** 3822–3830.

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

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

GENEST, C. and NEŠLEHOVÁ, J. G. (2014). Modeling dependence beyond correlation. In *Statistics in Action* 59–78. CRC Press, Boca Raton, FL. MR3241968

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

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.

GRIDELLI, C., MORGILLO, F., FAVARETTO, A., DE MARINIS, F., CHELLA, A., CEREA, G., MATTIOLI, R., TORTORA, G., ROSSI, A. et al. (2011). Sorafenib in combination with erlotinib or with gemcitabine in elderly patients with advanced non-small-cell lung cancer: A randomized phase II study. *Ann. Oncol.* **22** 1528–1534. https://doi.org/10.1093/annonc/mdq630

HA, M. J., STINGO, F. C. and BALADANDAYUTHAPANI, V. (2021). Bayesian structure learning in multilayered genomic networks. *J. Amer. Statist. Assoc.* **116** 605–618. MR4270007 https://doi.org/10.1080/01621459.2020.1775611

HO, C.-C., KUO, S.-H., HUANG, P.-H., HUANG, H.-Y., YANG, C.-H. and YANG, P.-C. (2008). Caveolin-1 expression is significantly associated with drug resistance and poor prognosis in advanced non-small cell lung cancer patients treated with gemcitabine-based chemotherapy. *Lung Cancer* **59** 105–110.

IORIO, F., KNIJNENBURG, T. A., VIS, D. J., BIGNELL, G. R., MENDEN, M. P., SCHUBERT, M., ABEN, N., GONÇALVES, E., BARTHORPE, S. et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell* **166** 740–754.

KASARSKIS, A., YANG, X. and SCHADT, E. (2011). Integrative genomics strategies to elucidate the complexity of drug response. *Pharmacogenomics J.* **12** 1695–1715. https://doi.org/10.2217/pgs.11.115

LANGER, C. J., LEIGHTON, J. C., COMIS, R. L., O'DWYER, P. J., MCALEER, C. A., BONJO, C. A., ENGSTROM, P. F., LITWIN, S. and OZOLS, R. F. (1995). Paclitaxel and carboplatin in combination in the treatment of advanced non-small-cell lung cancer: A phase II toxicity, response, and survival analysis. *J. Clin. Oncol.* **13** 1860–1870. https://doi.org/10.1200/JCO.1995.13.8.1860

LAURITZEN, S. L. and WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57. MR0981437 https://doi.org/10.1214/aos/1176347003

LI, Y., DATTA, J., CRAIG, B. A. and BHADRA, A. (2021). Joint mean-covariance estimation via the horseshoe. *J. Multivariate Anal.* **183** Paper No. 104716, 13. MR4196585 https://doi.org/10.1016/j.jmva.2020.104716

LIANG, H., WANG, H. B., LIU, H. Z., WEN, X. J., ZHOU, Q. L. and YANG, C. X. (2013). The effects of combined treatment with sevoflurane and cisplatin on growth and invasion of human adenocarcinoma cell line A549. *Biomed. Pharmacother.* **67** 503–509.

LIM, Z.-F. and MA, P. C. (2019). Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *J. Clin. Hematol. Oncol.* **12** 1–18.

LIN, J., BASU, S., BANERJEE, M. and MICHAILIDIS, G. (2016). Penalized maximum likelihood estimation of multi-layered Gaussian graphical models. *J. Mach. Learn. Res.* **17** Paper No. 146, 51. MR3555037

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084 https://doi.org/10.1214/12-AOS1037

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

MCCARTER, C. and KIM, S. (2014). On sparse Gaussian chain graph models. *Adv. Neural Inf. Process. Syst.* **27** 3212–3220.

MOK, T. S., WU, Y.-L., AHN, M.-J., GARASSINO, M. C., KIM, H. R., RAMALINGAM, S. S., SHEPHERD, F. A., HE, Y., AKAMATSU, H. et al. (2017). Osimertinib or platinum-pemetrexed in EGFR T790M-positive lung cancer. *N. Engl. J. Med.* **376** 629–640. https://doi.org/10.1056/NEJMoa1612674

MORRIS, J. S. and BALADANDAYUTHAPANI, V. (2017). Statistical contributions to bioinformatics: Design, modelling, structure learning and integration. *Stat. Model.* **17** 245–289. MR3706900 https://doi.org/10.1177/1471082X17698255

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x

PETERSEN, L. (2018). Sparse learning in Gaussian chain graphs for state space models. In *International Conference on Probabilistic Graphical Models* 332–343. PMLR.

PITT, M., CHAN, D. and KOHN, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93** 537–554. MR2261441 https://doi.org/10.1093/biomet/93.3.537

POMMIER, Y., SORDET, O., RAO, A., ZHANG, H. and KOHN, K. W. (2005). Targeting chk2 kinase: Molecular interaction maps and therapeutic rationale. *Curr. Pharm. Des.* **11** 2855.

ROBICHAUX, J. P., LE, X., VIJAYAN, R., HICKS, J. K., HEEKE, S., ELAMIN, Y. Y., LIN, H. Y., UDAGAWA, H., SKOULIDIS, F. et al. (2021). Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* **597** 732–737.

RODEN, D. M., MCLEOD, H. L., RELLING, M. V., WILLIAMS, M. S., MENSAH, G. A., PETERSON, J. F. and DRIEST, S. L. V. (2019). Pharmacogenomics. *Lancet* **394** 521–532. https://doi.org/10.1016/s0140-6736(19)31276-0

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19** 947–962. Supplementary materials available online. MR2791263 https://doi.org/10.1198/jcgs.2010.09188

ROVERATO, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat.* **29** 391–411. MR1925566 https://doi.org/10.1111/1467-9469.00297

SCHNEIDER, J., CLASSEN, V. and HELMIG, S. (2008). XRCC1 polymorphism and lung cancer risk. *Expert Rev. Mol. Diagn.* **8** 761–780. https://doi.org/10.1586/14737159.8.6.761

SHEN, Y., SOLÍS-LEMUS, C. and DESHPANDE, S. K. (2022). Sparse Gaussian chain graphs with the spike-and-slab LASSO: Algorithms and asymptotics. arXiv preprint. Available at arXiv:2207.07020.

SHI, Y.-B., LI, J., LAI, X.-N., JIANG, R., ZHAO, R.-C. and XIONG, L.-X. (2020). Multifaceted roles of caveolin-1 in lung cancer: A new investigation focused on tumor occurrence, development and therapy. *Cancers* **12** 291.

SORIA, J.-C., OHE, Y., VANSTEENKISTE, J., REUNGWETWATTANA, T., CHEWASKULYONG, B., LEE, K. H., DECHAPHUNKUL, A., IMAMURA, F., NOGAMI, N. et al. (2018). Osimertinib in untreated EGFR-mutated advanced non–small-cell lung cancer. *N. Engl. J. Med.* **378** 113–125.

SQUASSINA, A., MANCHIA, M., MANOLOPOULOS, V. G., ARTAC, M., LAPPA-MANAKOU, C., KARKABOUNA, S., MITROPOULOS, K., ZOMPO, M. D. and PATRINOS, G. P. (2010). Realities and expectations of pharmacogenomics and personalized medicine: Impact of translating genetic knowledge into clinical practice. *Pharmacogenomics J.* **11** 1149–1167.

VICENT, S., GARAYOA, M., LÓPEZ-PICAZO, J. M., LOZANO, M. D., TOLEDO, G., THUNNISSEN, F. B. J. M., MANZANO, R. G. and MONTUENGA, L. M. (2004). Mitogen-activated protein kinase phosphatase-1 is overexpressed in non-small cell lung cancer and is an independent predictor of outcome in patients. *Clin. Cancer Res.* **10** 3639–3649. https://doi.org/10.1158/1078-0432.CCR-03-0771

VOGELSTEIN, B. and KINZLER, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* **10** 789–799. https://doi.org/10.1038/nm1087

WOO, X. Y., GIORDANO, J., SRIVASTAVA, A., ZHAO, Z.-M., LLOYD, M. W., DE BRUIJN, R., SUH, Y.-S., PATIDAR, R., CHEN, L. et al. (2021). Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nat. Genet.* **53** 86–99. https://doi.org/10.1038/s41588-020-00750-6

XUE, L. and ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40** 2541–2571. MR3097612 https://doi.org/10.1214/12-AOS1041

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129 https://doi.org/10.1214/11-AOAS494

ZHOU, B.-B. S. and SAUSVILLE, E. A. (2003). Drug discovery targeting Chk1 and Chk2 kinases. *Prog. Cell Cycle Res.* **5** 413–421.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3297–3317
https://doi.org/10.1214/24-AOAS1937

A ROBUST BAYESIAN META-ANALYSIS FOR ESTIMATING THE HUBBLE CONSTANT VIA TIME DELAY COSMOGRAPHY

BY HYUNGSUK TAK[1,a] AND XUHENG DING[2,b]

[1]*Department of Statistics, The Pennsylvania State University,* [a]*tak@psu.edu*
[2]*School of Physics and Technology, Wuhan University,* [b]*dingxh@whu.edu.cn*

We propose a Bayesian meta-analysis to infer the current expansion rate of the Universe, called the Hubble constant (H_0), via time delay cosmography. Inputs of the meta-analysis are estimates of two properties for each pair of gravitationally lensed images; time delay and Fermat potential difference estimates with their standard errors. A meta-analysis can be appealing in practice because obtaining each estimate from even a single lens system involves substantial human efforts, and thus estimates are often separately obtained and published. Moreover, numerous estimates are expected to be available once the Rubin Observatory starts monitoring thousands of strong gravitational lens systems. This work focuses on combining these estimates from independent studies to infer H_0 in a robust manner. The robustness is crucial because currently up to eight lens systems are used to infer H_0, and thus any biased input can severely affect the resulting H_0 estimate. For this purpose we adopt Student's t error for the input estimates. We investigate properties of the resulting H_0 estimate via two simulation studies with realistic imaging data. It turns out that the meta-analysis can infer H_0 with sub-percent bias and about 1% level of coefficient of variation, even when 30% of inputs are manipulated to be outliers. We also apply the meta-analysis to three gravitationally lensed systems to obtain an H_0 estimate and compare it with existing estimates. An R package `h0` is publicly available for fitting the proposed meta-analysis.

REFERENCES

ABDALLA, E., ABELLÁN, G. F., ABOUBRAHIM, A. et al. (2022). Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies. *J. High Energy Astrophys*. **34** 49–211.

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

BIRRER, S. and AMARA, A. (2018). Lenstronomy: Multi-purpose gravitational lens modelling software package. *Phys. Dark Univers.* **22** 189–201.

BIRRER, S., AMARA, A. and REFREGIER, A. (2016). The mass-sheet degeneracy and time-delay cosmography: Analysis of the strong lens RXJ1131-1231. *J. Cosmol. Astropart. Phys.* **2016** 020.

BIRRER, S., MILLON, M., SLUSE, D. et al. (2022). Time-Delay Cosmography: Measuring the Hubble Constant and other cosmological parameters with strong gravitational lensing. ArXiv e-prints. Available at arXiv:2210.10833. https://doi.org/10.1007/s11214-024-01079-w

BIRRER, S., SHAJIB, A. J., GALAN, A. et al. (2020). TDCOSMO. IV. Hierarchical time-delay cosmography – joint inference of the Hubble constant and galaxy density profiles. *Astron. Astrophys.* **643** A165.

BIRRER, S., SHAJIB, A. J., GILMAN, D. et al. (2021). Lenstronomy II: A gravitational lensing software ecosystem. *J. Open Sour. Softw.* **6** 3283.

BIRRER, S., TREU, T., RUSU, C. E. et al. (2019). H0LiCOW – IX. Cosmographic analysis of the doubly imaged quasar SDSS 1206+4332 and a new measurement of the Hubble constant. *Mon. Not. R. Astron. Soc.* **484** 4726–4753.

BONVIN, V., CHAN, J. H. H., MILLON, M. et al. (2018). COSMOGRAIL—XVII. Time delays for the quadruply imaged quasar PG 1115+080. *Astron. Astrophys.* **616** A183.

Key words and phrases. Cosmology, Fermat potential, gravitational lensing, hierarchical model.

Bonvin, V., Courbin, F., Suyu, S. H. et al. (2016). H0LiCOW V. New COSMOGRAIL time delays of HE 0435–1223: H_0 to 3.8 per cent precision from strong lensing in a flat ΛCDM model. *Mon. Not. R. Astron. Soc.* **465** 4914–4930.

Bonvin, V., Millon, M., Chan, J. H. H. et al. (2019). COSMOGRAIL—XVIII. Time delays of the quadruply lensed quasar WFI2033-4723. *Astron. Astrophys.* **629** A97.

Buckley-Geer, E. J., Lin, H., Rusu, C. E. et al. (2020). STRIDES: Spectroscopic and photometric characterization of the environment and effects of mass along the line of sight to the gravitational lenses DES J0408-5354 and WGD 2038-4008. *Mon. Not. R. Astron. Soc.* **498** 3241–3274.

Chen, G. C. F., Fassnacht, C. D., Suyu, S. H. et al. (2019). A SHARP view of H0LiCOW: H_0 from three time-delay gravitational lens systems with adaptive optics imaging. *Mon. Not. R. Astron. Soc.* **490** 1743–1773.

Chen, G. C. F., Fassnacht, C. D., Suyu, S. H. et al. (2021). TDCOSMO. VI. Distance measurements in time-delay cosmography under the mass-sheet transformation. *Astron. Astrophys.* **652** A7.

Courbin, F., Bonvin, V., Buckley-Geer, E. et al. (2018). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses—XVI. Time delays for the quadruply imaged quasar DES J0408-5354 with high-cadence photometric monitoring. *Astron. Astrophys.* **609** A71.

Cowan, G. (2019). Statistical models with uncertain error parameters. *European Physical Journal C* **79** 133.

Denzel, P., Coles, J. P., Saha, P. and Williams, L. L. R. (2021). The Hubble constant from eight time-delay galaxy lenses. *Mon. Not. R. Astron. Soc.* **501** 784–801.

Di Valentino, E., Mena, O., Pan, S. et al. (2021). In the realm of the Hubble tension-a review of solutions. *Classical Quantum Gravity* **38** 153001.

Ding, X., Liao, K., Birrer, S. et al. (2021a). Improved time-delay lens modelling and H_0 inference with transient sources. *Mon. Not. R. Astron. Soc.* **504** 5621–5628.

Ding, X., Treu, T., Birrer et al. (2021b). Time delay lens modelling challenge. *Mon. Not. R. Astron. Soc.* **503** 1096–1123.

Eigenbrod, A., Courbin, F., Vuissoz, C. et al. (2005). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses. I. How to sample the light curves of gravitationally lensed quasars to measure accurate time delays. *Astron. Astrophys.* **436** 25–35.

Ertl, S., Schuldt, S., Suyu, S. H. et al. (2023). TDCOSMO. X. Automated modeling of nine strongly lensed quasars and comparison between lens-modeling software. *Astron. Astrophys.* **672** A2.

Eulaers, E., Tewes, M., Magain, P. et al. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses. XII. Time delays of the doubly lensed quasars SDSS J1206+4332 and HS 2209+1914. *Astron. Astrophys.* **553** A121.

Fleury, P., Larena, J. and Uzan, J.-P. (2021). Line-of-sight effects in strong gravitational lensing. *J. Cosmol. Astropart. Phys.* **8** Paper No. 024. MR4348503 https://doi.org/10.1088/1475-7516/2021/08/024

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.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741. https://doi.org/10.1109/tpami.1984.4767596

Gorenstein, M. V., Falco, E. E. and Shapiro, I. I. (1988). Degeneracies in parameter estimates for models of gravitational lens systems. *Astrophys. J.* **327** 693.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91** 1461–1473. MR1439086 https://doi.org/10.2307/2291572

Hogg, D. W. (1999). Distance Measures in Cosmology. ArXiv preprint. Available at arXiv:astro-ph/9905116.

Hu, Z. and Tak, H. (2020). Modeling stochastic variability in multiband time-series data. *Astron. J.* **160** 265.

Kelly, B. C., Bechtold, J. and Siemiginowska, A. (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *Astrophys. J.* **698** 895.

Leon-Anaya, L., Cuevas-Tello, J. C., Valenzuela, O. et al. (2023). Data science methodology for time-delay estimation and data preprocessing of the time-delay challenge. *Mon. Not. R. Astron. Soc.* **522** 1323–1341.

Liao, K., Treu, T., Marshall, P. et al. (2015). Strong lens time delay challenge. II. Results of TDC1. *Astrophys. J.* **800** 11.

Linder, E. V. (2011). Lensing time delays and cosmological complementarity. *Phys. Rev. D* **84** 123529.

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

Meyer, A. D., van Dyk, D. A., Tak, H. and Siemiginowska, A. (2023). TD-CARMA: Painless, accurate, and scalable estimates of gravitational lens time delays with flexible CARMA processes. *Astrophys. J.* **950** 37.

Millon, M., Courbin, F., Bonvin, V. et al. (2020a). TDCOSMO. II. Six new time delays in lensed quasars from high-cadence monitoring at the MPIA 2.2 m telescope. *Astron. Astrophys.* **642** A193.

MILLON, M., COURBIN, F., BONVIN, V. et al. (2020b). COSMOGRAIL. XIX. Time delays in 18 strongly lensed quasars from 15 years of optical monitoring. *Astron. Astrophys.* **640** A105.

MILLON, M., GALAN, A., COURBIN, F. et al. (2020c). TDCOSMO—I. An exploration of systematic uncertainties in the inference of H0 from time-delay cosmography. *Astron. Astrophys.* **639** A101.

OGURI, M. and MARSHALL, P. J. (2010). Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. *Mon. Not. R. Astron. Soc.* **405** 2579–2593.

PLANCK COLLABORATION AGHANIM, N., AKRAMI, Y., ASHDOWN, M. et al. (2020). Planck 2018 results—VI. Cosmological parameters. *Astron. Astrophys.* **641** A6.

REFSDAL, S. (1964). On the possibility of determining Hubble's parameter and the masses of galaxies from the gravitational lens effect. *Mon. Not. R. Astron. Soc.* **128** 307–310. MR0175608 https://doi.org/10.1093/mnras/128.4.307

RIESS, A. G., CASERTANO, S., YUAN, W. et al. (2019). Large magellanic cloud Cepheid standards provide a 1% foundation for the determination of the Hubble constant and stronger evidence for physics beyond ΛCDM. *Astrophys. J.* **876** 85.

RIESS, A. G., CASERTANO, S., YUAN, W. et al. (2021). Cosmic distances calibrated to 1% precision with gaia EDR3 parallaxes and Hubble space telescope photometry of 75 Milky Way cepheids confirm tension with ΛCDM. *Astrophys. J.* **908** L6.

RIESS, A. G., YUAN, W., MACRI, L. M. et al. (2022). A comprehensive measurement of the local value of the Hubble constant with 1 km s^{-1} Mpc^{-1} uncertainty from the Hubble space telescope and the SH0ES team. *Astrophys. J.* **934** L7.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

RUSU, C. E., FASSNACHT, C. D., SLUSE, D. et al. (2017). H0LiCOW—III. Quantifying the effect of mass along the line of sight to the gravitational lens HE 0435-1223 through weighted galaxy counts. *Mon. Not. R. Astron. Soc.* **467** 4220–4242.

RUSU, C. E., WONG, K. C., BONVIN, V. et al. (2020). H0LiCOW XII. Lens mass model of WFI2033-4723 and blind measurement of its time-delay distance and H_0. *Mon. Not. R. Astron. Soc.* **498** 1440–1468.

SCHMIDT, T., TREU, T., BIRRER, S. et al. (2023). STRIDES: Automated uniform models for 30 quadruply imaged quasars. *Mon. Not. R. Astron. Soc.* **518** 1260–1300.

SCHNEIDER, P. and SLUSE, D. (2013). Mass-sheet degeneracy, power-law models and external convergence: Impact on the determination of the Hubble constant from gravitational lensing. *Astron. Astrophys.* **559** A37.

SCHNEIDER, P., WAMBSGANSS, J. and KOCHANEK, C. S. (2006). *Gravitational Lensing*: *Strong*, *Weak and Micro*. Springer, Berlin.

SERENO, M. and PARAFICZ, D. (2014). Hubble constant and dark energy inferred from free-form determined time delay distances. *Mon. Not. R. Astron. Soc.* **437** 600–605.

SHAH, P., LEMOS, P. and LAHAV, O. (2021). A buyer's guide to the Hubble constant. *Astron. Astrophys. Rev.* **29** 9.

SHAJIB, A. J., BIRRER, S., TREU, T. et al. (2019). Is every strong lens model unhappy in its own way? Uniform modelling of a sample of 13 quadruply+ imaged quasars. *Mon. Not. R. Astron. Soc.* **483** 5649–5671.

SHAJIB, A. J., BIRRER, S., TREU, T. et al. (2020). STRIDES: A 3.9 per cent measurement of the Hubble constant from the strong lens system DES J0408-5354. *Mon. Not. R. Astron. Soc.* **494** 6072–6102.

SHAJIB, A. J., MOZUMDAR, P., CHEN, G. C. F. et al. (2023). TDCOSMO. XII. Improved Hubble constant measurement from lensing time delays using spatially resolved stellar kinematics of the lens galaxy. *Astron. Astrophys.* **673** A9.

SHAJIB, A. J., WONG, K. C., BIRRER, S. et al. (2022). TDCOSMO. IX. Systematic comparison between lens modelling software programs: Time delay prediction for WGD 2038–4008. ArXiv e-prints. Available at arXiv:2202.11101. https://doi.org/10.1051/0004-6361/202243401

SHALYAPIN, V. N., GOICOECHEA, L. J. and GIL-MERINO, R. (2012). A 5.5-year robotic optical monitoring of Q0957+561: Substructure in a non-local cD galaxy. *Astron. Astrophys.* **540** A132.

SUYU, S. H., AUGER, M. W., HILBERT, S. et al. (2013). Two accurate time-delay distances from strong lensing: Implications for cosmology. *Astrophys. J.* **766** 70.

SUYU, S. H., BONVIN, V., COURBIN, F. et al. (2017). H0LiCOW I. H0 lenses in COSMOGRAIL's wellspring: Program overview. *Mon. Not. R. Astron. Soc.* **468** 2590–2604.

SUYU, S. H., MARSHALL, P. J., AUGER, M. W. et al. (2010). Dissecting the gravitational lens B1608+656. II. Precision measurements of the Hubble constant spatial curvature, and the dark energy equation of state. *Astrophys. J.* **711** 201–221.

SUYU, S. H., TREU, T., HILBERT, S. et al. (2014). Cosmology from gravitational lens time delays and Planck data. *Astrophys. J.* **788** L35.

TAK, H. and DING, X. (2024). Supplement to "A robust Bayesian meta-analysis for estimating the Hubble constant via time delay cosmography." https://doi.org/10.1214/24-AOAS1937SUPPA, https://doi.org/10.1214/24-AOAS1937SUPPB

TAK, H., ELLIS, J. A. and GHOSH, S. K. (2019). Robust and accurate inference via a mixture of Gaussian and Student's t errors. *J. Comput. Graph. Statist.* **28** 415–426. MR3974890 https://doi.org/10.1080/10618600.2018.1537925

TAK, H., GHOSH, S. K. and ELLIS, J. A. (2018). How proper are Bayesian models in the astronomical literature? *Mon. Not. R. Astron. Soc.* **481** 277–285.

TAK, H., MANDEL, K., VAN DYK, D. A., KASHYAP, V. L., MENG, X.-L. and SIEMIGINOWSKA, A. (2017). Bayesian estimates of astronomical time delays between gravitationally lensed stochastic light curves. *Ann. Appl. Stat.* **11** 1309–1348. MR3709561 https://doi.org/10.1214/17-AOAS1027

TAK, H., MENG, X.-L. and VAN DYK, D. A. (2018). A repelling-attracting Metropolis algorithm for multimodality. *J. Comput. Graph. Statist.* **27** 479–490. MR3863751 https://doi.org/10.1080/10618600.2017.1415911

TEWES, M., COURBIN, F., MEYLAN, G. et al. (2013). COSMOGRAIL: The COSmological MOnitoring of GRAvItational lenses XIII: Time delays and 9-yr optical monitoring of the lensed quasar RX J1131-1231. *Astron. Astrophys.* **556** A22.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166 https://doi.org/10.1214/aos/1176325750

TIHHONOVA, O., COURBIN, F., HARVEY, D. et al. (2020). H0LiCOW—XI. A weak lensing measurement of the external convergence in the field of the lensed quasar B1608+656 using HST and Subaru deep imaging. *Mon. Not. R. Astron. Soc.* **498** 1406–1419.

TREU, T. (2010). Strong lensing by galaxies. *Annu. Rev. Astron. Astrophys.* **48** 87–125.

TREU, T., AGNELLO, A., BAUMER, M. A. et al. (2018). The STRong lensing insights into the dark energy survey (STRIDES) 2016 follow-up campaign—I. Overview and classification of candidates selected by two techniques. *Mon. Not. R. Astron. Soc.* **481** 1041–1054.

TREU, T. and MARSHALL, P. J. (2016). Time delay cosmography. *Astron. Astrophys.* **24** 11.

TREU, T., SUYU, S. H. and MARSHALL, P. J. (2022). Strong lensing time-delay cosmography in the 2020s. *Astron. Astrophys. Rev.* **30** 8.

VERDE, L., TREU, T. and RIESS, A. G. (2019). Tensions between the early and the late universe. *Nat. Astron.* **3** 891–895.

WANG, L.-F., ZHANG, J.-H., HE, D.-Z. et al. (2022). Constraints on interacting dark energy models from time-delay cosmography with seven lensed quasars. *Mon. Not. R. Astron. Soc.* **514** 1433–1440.

WELLS, P., FASSNACHT, C. D. and RUSU, C. E. (2023). TDCOSMO XIV: Practical Techniques for Estimating External Convergence of Strong Gravitational Lens Systems and Applications to the SDSS J0924+0219 System. ArXiv e-prints. Available at arXiv:2302.03176. https://doi.org/10.1051/0004-6361/202346093

WONG, K. C., SUYU, S. H., AUGER, M. W. et al. (2017). H0LiCOW—IV. Lens mass model of HE 0435-1223 and blind measurement of its time-delay distance for cosmology. *Mon. Not. R. Astron. Soc.* **465** 4895–4913.

WONG, K. C., SUYU, S. H., CHEN, G. C. F. et al. (2020). H0LiCOW XIII. A 2.4% measurement of H_0 from lensed quasars: 5.3σ tension between early and late-Universe probes. *Mon. Not. R. Astron. Soc.* **498** 1420–1439.

YILDIRIM, A., SUYU, S. H., CHEN, G. C. F. and KOMATSU, E. (2023). TDCOSMO. XIII. Cosmological distance measurements in light of the mass-sheet degeneracy: Forecasts from strong lensing and integral field unit stellar kinematics. *Astron. Astrophys.* **675** A21.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3318–3337
https://doi.org/10.1214/24-AOAS1938

A SEMIPARAMETRIC METHOD FOR RISK PREDICTION USING INTEGRATED ELECTRONIC HEALTH RECORD DATA

BY JILL HASLER[1,a], YANYUAN MA[2,b], YIZHENG WEI[3,c], RAVI PARIKH[4,d] AND JINBO CHEN[5,e]

[1]*Fox Chase Cancer Center,* [a]*Jill.Hasler@fccc.edu*

[2]*Department of Statistics, Pennsylvania State University,* [b]*yanyuanma@gmail.com*

[3]*Department of Statistics, University of South Carolina,* [c]*wyz.lewis@gmail.com*

[4]*Departments of Medical Ethics and Health Policy and Medicine, University of Pennsylvania,* [d]*ravi.parikh@pennmedicine.upenn.edu*

[5]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,* [e]*jinboche@pennmedicine.upenn.edu*

When using electronic health records (EHRs) for clinical and translational research, additional data is often available from external sources to enrich the information extracted from EHRs. For example, academic biobanks have more granular data available, and patient reported data is often collected through small-scale surveys. It is common that the external data is available only for a small subset of patients who have EHR information. We propose efficient and robust methods for building and evaluating models for predicting the risk of binary outcomes using such integrated EHR data. Our method is built upon an idea derived from the two-phase design literature that modeling the availability of a patient's external data as a function of an EHR-based preliminary predictive score leads to effective utilization of the EHR data. Through both theoretical and simulation studies, we show that our method has high efficiency for estimating log-odds ratio parameters, the area under the ROC curve, as well as other measures for quantifying predictive accuracy. We apply our method to develop a model for predicting the short-term mortality risk of oncology patients, where the data was extracted from the University of Pennsylvania hospital system EHR and combined with survey-based patient reported outcome data.

REFERENCES

AMORIM, G., TAO, R., LOTSPEICH, S., SHAW, P. A., LUMLEY, T. and SHEPHERD, B. E. (2021). Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. *J. Roy. Statist. Soc. Ser. A* **184** 1368–1389. MR4344641 https://doi.org/10.1111/rssa.12689

BASCH, E., BARBERA, L., KERRIGAN, C. L. and VELIKOVA, G. (2018). Implementation of patient-reported outcomes in routine medical care. *Amer. Soc. Clin. Oncol. Educ. Book* **38** 122–134. https://doi.org/10.1200/EDBK_200383

BASCH, E., DEAL, A. M., KRIS, M. G., SCHER, H. I., HUDIS, C. A., SABBATINI, P., ROGAK, L., BENNETT, A. V., DUECK, A. C. et al. (2016). Symptom monitoring with patient-reported outcomes during routine cancer treatment: A randomized controlled trial. *J. Clin. Oncol.* **34** 557–565. https://doi.org/10.1200/JCO.2015.63.0830

BORGAN, Ø. and SAMUELSEN, S. O. (2014). Nested case-control and case-cohort studies. In *Handbook of Survival Analysis. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 343–367. CRC Press, Boca Raton, FL. MR3287454

BRESLOW, N., MCNENEY, B. and WELLNER, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* **31** 1110–1139. MR2001644 https://doi.org/10.1214/aos/1059655907

BRESLOW, N. E. and CAIN, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75** 11–20. MR0932812 https://doi.org/10.1093/biomet/75.1.11

BRESLOW, N. E. and CHATTERJEE, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **48** 457–468.

Key words and phrases. Area under the ROC curve (AUC), integrated EHR data, semiparametric estimation, two-phase design.

BRESLOW, N. E. and HOLUBKOV, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc. Ser. B* **59** 447–461. MR1440590 https://doi.org/10.1111/1467-9868.00078

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009). Using the whole cohort in the analysis of case-cohort data. *Amer. J. Epidemiol.* **169** 1398–1405. https://doi.org/10.1093/aje/kwp055

CAI, T. and ZHENG, Y. (2012). Evaluating prognostic accuracy of biomarkers in nested case-control studies. *Biostatistics* **13** 89–100.

CAI, T. and ZHENG, Y. (2013). Resampling procedures for making inference under nested case-control studies. *J. Amer. Statist. Assoc.* **108** 1532–1544. MR3174727 https://doi.org/10.1080/01621459.2013.856715

CAO, Y., HANEUSE, S., ZHENG, Y. and CHEN, J. (2023). Two-phase stratified sampling and analysis for predicting binary outcomes. *Biostatistics* **24** 585–602. MR4615242 https://doi.org/10.1093/biostatistics/kxab044

CHE, M., HAN, P. and LAWLESS, J. F. (2023). Improving estimation efficiency for two-phase, outcome-dependent sampling studies. *Electron. J. Stat.* **17** 1043–1073. MR4571186 https://doi.org/10.1214/23-ejs2124

CHE, M., LAWLESS, J. F. and HAN, P. (2021). Empirical and conditional likelihoods for two-phase studies. *Canad. J. Statist.* **49** 344–361. MR4267924 https://doi.org/10.1002/cjs.11566

CHOUDHURY, P., CHATURVEDI, A. K. and CHATTERJEE, N. (2020). Evaluating discrimination of a lung cancer risk prediction model using partial risk-score in a two-phase study. *Cancer Epidemiol. Biomark. Prev.* **29** 1196–1203.

CHRISTAKIS, N. A., SMITH, J. L., PARKES, C. M. and LAMONT, E. B. (2000). Extent and determinants of error in doctors' prognoses in terminally ill patients: Prospective cohort study. Commentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition. *BMJ* **320** 469–473.

ELFIKY, A. A., PANY, M. J., PARIKH, R. B. and OBERMEYER, Z. (2018). Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Netw. Open* **1** e180926. https://doi.org/10.1001/jamanetworkopen.2018.0926

FITHIAN, W. and HASTIE, T. (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Ann. Statist.* **42** 1693–1724. MR3257627 https://doi.org/10.1214/14-AOS1220

FLANDERS, W. D. and GREENLAND, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Stat. Med.* **10** 739–747.

GENSHEIMER, M. F., HENRY, A. S., WOOD, D. J., HASTIE, T. J., AGGARWAL, S., DUDLEY, S. A., PRADHAN, P., BANERJEE, I., CHO, E. et al. (2019). Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *J. Natl. Cancer Inst.* **111** 568–574.

GOLDSTEIN, B. A., NAVAR, A. M., PENCINA, M. J. and IOANNIDIS, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Amer. Med. Inform. Assoc.* **24** 198–208. https://doi.org/10.1093/jamia/ocw042

GRAMLING, R., GAJARY-COOTS, E., CIMINO, J., FISCELLA, K., EPSTEIN, R., LADWIG, S., ANDERSON, W., ALEXANDER, S. C., HAN, P. K. et al. (2019). Palliative care clinician overestimation of survival in advanced cancer: Disparities and association with end-of-life care. *J. Pain Symptom Manag.* **57** 233–240.

HASLER, J., MA, Y., WEI, Y., PARIKH, R. and CHEN, J. (2024). Supplement to "A Semiparametric Method for Risk Prediction Using Integrated Electronic Health Record Data." https://doi.org/10.1214/24-AOAS1938SUPPA, https://doi.org/10.1214/24-AOAS1938SUPPB

HENMI, M. and EGUCHI, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91** 929–941. MR2126042 https://doi.org/10.1093/biomet/91.4.929

HUANG, Y. (2016). Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case-control studies. *Biostatistics* **17** 499–522. MR3603950 https://doi.org/10.1093/biostatistics/kxw003

HUANG, Y. and PEPE, M. S. (2010). Assessing risk prediction models in case-control studies using semiparametric and nonparametric methods. *Stat. Med.* **29** 1391–1410. MR2758124 https://doi.org/10.1002/sim.3876

LAWLESS, J. F., KALBFLEISCH, J. D. and WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 413–438. MR1680310 https://doi.org/10.1111/1467-9868.00185

LIN, D.-Y., ZENG, D. and TANG, Z.-Z. (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proc. Natl. Acad. Sci. USA* **110** 12247–12252. MR3105371 https://doi.org/10.1073/pnas.1221713110

LIPSITZ, S. R., IBRAHIM, J. G. and ZHAO, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc.* **94** 1147–1160. MR1731479 https://doi.org/10.2307/2669931

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0890519

LIU, X. and ZHAO, Y. (2012). Semi-empirical likelihood inference for the ROC curve with missing data. *J. Statist. Plann. Inference* **142** 3123–3133. MR2956798 https://doi.org/10.1016/j.jspi.2012.06.011

LONG, Q., ZHANG, X. and HSU, C.-H. (2011). Nonparametric multiple imputation for receiver operating characteristics analysis when some biomarker values are missing at random. *Stat. Med.* **30** 3149–3161. MR2845684 https://doi.org/10.1002/sim.4338

LUMLEY, T. (2011). *Complex Surveys*: *A Guide to Analysis Using R*. Wiley, New York.

MANZ, C. R., PARIKH, R. B., SMALL, D. S., EVANS, C. N., CHIVERS, C., REGLI, S. H., HANSON, C. W., BEKELMAN, J. E., RARESHIDE, C. A. et al. (2020). Effect of integrating machine learning mortality estimates with behavioral nudges to clinicians on serious illness conversations among patients with cancer: A stepped-wedge cluster randomized clinical trial. *JAMA Oncol.* **6** e204759.

MANZ, C. R., ZHANG, Y., CHEN, K., LONG, Q., SMALL, D. S., EVANS, C. N., CHIVERS, C., REGLI, S. H., HANSON, C. W. et al. (2023). Long-term effect of machine learning–triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: A randomized clinical trial. *JAMA Oncol.* **9** 414–418.

MARONGE, J. M., TAO, R., SCHILDCROUT, J. S. and RATHOUZ, P. J. (2023). Generalized case-control sampling under generalized linear models. *Biometrics* **79** 332–343. MR4572525 https://doi.org/10.1111/biom.13571

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33** 101–116.

PARIKH, R. B., MANZ, C., CHIVERS, C., REGLI, S. H., BRAUN, J., DRAUGELIS, M. E., SCHUCHTER, L. M., SHULMAN, L. N., NAVATHE, A. S. et al. (2019). Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw. Open* **2** e1915997. https://doi.org/10.1001/jamanetworkopen.2019.15997

PAYNE, R., YANG, M., ZHENG, Y., JENSEN, M. K. and CAI, T. (2016). Robust risk prediction with biomarkers under two-phase stratified cohort design. *Biometrics* **72** 1037–1045. MR3591588 https://doi.org/10.1111/biom.12515

PIERCE, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Ann. Statist.* **10** 475–478. MR0653522

PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.

QIN, J., ZHANG, B. and LEUNG, D. H. Y. (2017). Efficient augmented inverse probability weighted estimation in missing data problems. *J. Bus. Econom. Statist.* **35** 86–97. MR3591539 https://doi.org/10.1080/07350015.2015.1058266

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

SCOTT, A. J. and WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84** 57–71. MR1450191 https://doi.org/10.1093/biomet/84.1.57

SCOTT, A. J. and WILD, C. J. (2011). Fitting regression models with response-biased samples. *Canad. J. Statist.* **39** 519–536. MR2842429 https://doi.org/10.1002/cjs.10114

SCOTTÉ, F., TAYLOR, A. and DAVIES, A. (2023). Supportive care: The "Keystone" of modern oncology practice. *Cancers* **15** 3860. https://doi.org/10.3390/cancers15153860

SHARMA, V., ALI, I., VEER, S. V. D., MARTIN, G., AINSWORTH, J. and AUGUSTINE, T. (2021). Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform.* **28** e100253. https://doi.org/10.1136/bmjhci-2020-100253

SONG, R., ZHOU, H. and KOSOROK, M. R. (2009). A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika* **96** 221–228. MR2482147 https://doi.org/10.1093/biomet/asn073

STEYERBERG, E. W. (2019). Validation of prediction models. In *Clinical Prediction Models*: *A Practical Approach to Development*, *Validation*, *and Updating* (E. W. Steyerberg, ed.). *Statistics for Biology and Health* 329–344. Springer, Cham.

TAN, W. K. and HEAGERTY, P. J. (2020). Predictive case control designs for modification learning. arXiv:2011.14529 [stat].

TAO, R., ZENG, D. and LIN, D.-Y. (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *J. Amer. Statist. Assoc.* **112** 1468–1476. MR3750869 https://doi.org/10.1080/01621459.2017.1295864

TAO, R., ZENG, D. and LIN, D.-Y. (2020). Optimal designs of two-phase studies. *J. Amer. Statist. Assoc.* **115** 1946–1959. MR4189769 https://doi.org/10.1080/01621459.2019.1671200

WANG, K., EFTANG, C. N., JAKOBSEN, R. B. and ÅRØEN, A. (2020a). Review of response rates over time in registry-based studies using patient-reported outcome measures. *BMJ* **10** e030808.

WANG, L. and HUANG, Y. (2019). Evaluating classification performance of biomarkers in two-phase case-control studies. *Stat. Med.* **38** 100–114. MR3887270 https://doi.org/10.1002/sim.7966

WANG, L., WILLIAMS, M. L., CHEN, Y. and CHEN, J. (2020b). Novel two-phase sampling designs for studying binary outcomes. *Biometrics* **76** 210–223. MR4098556 https://doi.org/10.1111/biom.13140

WEEKS, J. C., COOK, E. F., O'DAY, S. J., PETERSON, L. M., WENGER, N., REDING, D., HARRELL, F. E., KUSSIN, P., DAWSON, N. V. et al. (1998). Relationship between cancer patients' predictions of prognosis and their treatment preferences. *JAMA* **279** 1709–1714.

WHITE, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Amer. J. Epidemiol*. **115** 119–128.

YANG, L. Y., MANHAS, D. S., HOWARD, A. F. and OLSON, R. A. (2018). Patient-reported outcome use in oncology: A systematic review of the impact on patient-clinician communication. *Support. Care Cancer* **26** 41–60. https://doi.org/10.1007/s00520-017-3865-7

YAO, W., LI, Z. and GRAUBARD, B. I. (2015). Estimation of ROC curve with complex survey data. *Stat. Med*. **34** 1293–1303. MR3322769 https://doi.org/10.1002/sim.6405

YILMAZ, Y. E. and BULL, S. B. (2011). Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? *BMC Proc*. **5** S111. https://doi.org/10.1186/1753-6561-5-S9-S111

ZHENG, Y., BROWN, M., LOK, A. and CAI, T. (2017). Improving efficiency in biomarker incremental value evaluation under two-phase designs. *Ann. Appl. Stat*. **11** 638–654. MR3693540 https://doi.org/10.1214/16-AOAS997

ZHENG, Y., CAI, T. and PEPE, M. S. (2013). Adopting nested case-control quota sampling designs for the evaluation of risk markers. *Lifetime Data Anal*. **19** 568–588. MR3119997 https://doi.org/10.1007/s10985-013-9270-8

ZHOU, Q. M., ZHENG, Y., CHIBNIK, L. B., KARLSON, E. W. and CAI, T. (2015). Assessing incremental value of biomarkers with multi-phase nested case-control studies. *Biometrics* **71** 1139–1149. MR3436739 https://doi.org/10.1111/biom.12344

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3338–3363
https://doi.org/10.1214/24-AOAS1939

POISSON–BIRNBAUM–SAUNDERS REGRESSION MODEL FOR CLUSTERED COUNT DATA

BY JUSSIANE NADER GONÇALVES[1,a], WAGNER BARRETO-SOUZA[2,b] AND HERNANDO OMBAO[3,c]

[1]*Departamento de Estatística, Universidade Federal de Minas Gerais,* [a]*jussianegoncalves@gmail.com*
[2]*School of Mathematics and Statistics, University College Dublin,* [b]*wagner.barreto-souza@ucd.ie*
[3]*Statistics Program, King Abdullah University of Science and Technology,* [c]*hernando.ombao@kaust.edu.sa*

In this paper we study the number of inpatient admissions by individuals to hospital emergency rooms reported by the 2003 Medical Expenditure Panel Survey (MEPS), which the United States Agency for Health Research and Quality conducts. Explanatory variables such as health status, access, use, and costs of health services in the U.S.A. are considered. Our main goal is to properly model the number of inpatient admissions, according to the geographical U.S. regions, as a tool for measuring the volume of diagnostic procedures in the health care system. In the analysis four clusters were determined according to the regions in the U.S., namely, the midwest, northeast, south, and west. The clustered analysis of this count data from the MEPS is a novel contribution to the best of our knowledge. Our analysis demonstrated that a clustered negative binomial (CNB) regression (Poisson model with latent gamma effects) might not be a suitable choice for analyzing the MEPS data. This fact motivates us to introduce a new regression model to handle clustered count data. To account for correlation within clusters, we propose a Poisson regression model where the observations within the same cluster are driven by the same latent random effect that follows a Birnbaum–Saunders distribution with a parameter that controls the strength of dependence among the individuals. This novel multivariate count model is called Clustered Poisson–Birnbaum–Saunders (CPBS) regression. The CPBS model is analytically tractable, and its moment structure can be explicitly obtained. We also derive theoretical/methodological studies to advise when the Birnbaum–Saunders effect should be preferred over the gamma effect (and vice-versa) in terms of probability tail. Estimation is performed through the maximum likelihood method. Here we also developed an expectation-maximization (EM) algorithm for estimation. Simulation results that evaluate the finite-sample performance of our proposed estimators are presented. Studies on the potential impact of model misspecification were conducted, and comparisons between our model and a CNB regression were also addressed. A full statistical analysis of the MEPS data reveals that, compared to the CNB model, the CPBS regression model produces better results in terms of prediction and goodness-of-fit.

REFERENCES

ATKINSON, A. C. (1985). *Plots*, *Transformations*, *and Regression*. Oxford Univ. Press, Oxford.

BARRETO-SOUZA, W. and SIMAS, A. B. (2016). General mixed Poisson regression models with varying dispersion. *Stat. Comput.* **26** 1263–1280. MR3538636 https://doi.org/10.1007/s11222-015-9601-6

BASTOS, F. S. and BARRETO-SOUZA, W. (2021). Birnbaum–Saunders sample selection model. *J. Appl. Stat.* **48** 1896–1916. MR4296771 https://doi.org/10.1080/02664763.2020.1780570

BIRNBAUM, Z. W. and SAUNDERS, S. C. (1969). A new family of life distributions. *J. Appl. Probab.* **6** 319–327. MR0253493 https://doi.org/10.2307/3212003

Key words and phrases. Covariates, diagnostic tools, EM-algorithm, maximum likelihood estimation, multivariate Poisson–Birnbaum–Saunders distribution.

CAMERON, A. C. and TRIVEDI, P. K. (2013). *Regression Analysis of Count Data*, 2nd ed. *Econometric Society Monographs* **53**. Cambridge Univ. Press, Cambridge. MR3155491 https://doi.org/10.1017/CBO9781139013567

CHOO-WOSOBA, H. and DATTA, S. (2018). Analyzing clustered count data with a cluster-specific random effect zero-inflated Conway–Maxwell–Poisson distribution. *J. Appl. Stat.* **45** 799–814. MR3772111 https://doi.org/10.1080/02664763.2017.1312299

CHOO-WOSOBA, H., GASKINS, J., LEVY, S. and DATTA, S. (2018). A Bayesian approach for analyzing zero-inflated clustered count data with dispersion. *Stat. Med.* **37** 801–812. MR3760450 https://doi.org/10.1002/sim.7541

CHOO-WOSOBA, H., LEVY, S. M. and DATTA, S. (2016). Marginal regression models for clustered count data based on zero-inflated Conway–Maxwell–Poisson distribution with applications. *Biometrics* **72** 606–618. MR3515787 https://doi.org/10.1111/biom.12436

CONSUL, P. C. and FAMOYE, F. (1992). Generalized Poisson regression model. *Comm. Statist. Theory Methods* **21** 89–109.

COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19** 15–18. MR0436478 https://doi.org/10.2307/1268249

DEAN, C., LAWLESS, J. F. and WILLMOT, G. E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canad. J. Statist.* **17** 171–181. MR1033100 https://doi.org/10.2307/3314846

DEMIDENKO, E. (2007). Poisson regression for clustered data. *Int. Stat. Rev.* **75** 96–113.

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

DIMITRIS, K. and XEKALAKI, E. (2005). Mixed Poisson distributions. *Int. Stat. Rev.* **73** 35–58.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

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

FABIO, L. C., PAULA, G. A. and DE CASTRO, M. (2012). A Poisson mixed model with nonnormal random effect distribution. *Comput. Statist. Data Anal.* **56** 1499–1510. MR2892356 https://doi.org/10.1016/j.csda.2011.12.002

FAMOYE, F. and SINGH, K. P. (2004). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *J. Data Sci.* **4** 117–130.

FREES, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. *International Series on Actuarial Science*. Cambridge Univ. Press, Cambridge. MR2572259

GÓMEZ-DÉNIZ, E., GHITANY, M. E. and GUPTA, R. C. (2016). Poisson-mixed inverse Gaussian regression model and its application. *Comm. Statist. Simulation Comput.* **45** 2767–2781. MR3514839 https://doi.org/10.1080/03610918.2014.925924

GONÇALVES, J. N. and BARRETO-SOUZA, W. (2020). Flexible regression models for counts with high-inflation of zeros. *Metron* **78** 71–95. MR4081358 https://doi.org/10.1007/s40300-020-00163-9

GONÇALVES, J. N., BARRETO-SOUZA, W. and OMBAO, H. (2024). Supplement to "Poisson–Birnbaum–Saunders regression model for clustered count data." https://doi.org/10.1214/24-AOAS1939SUPPA, https://doi.org/10.1214/24-AOAS1939SUPPB

GUO, G. (1996). Negative multinomial regression models for clustered event counts. *Sociol. Method.* **26** 113–132.

HALL, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56** 1030–1039. MR1815581 https://doi.org/10.1111/j.0006-341X.2000.01030.x

HALL, D. B. and ZHANG, Z. (2004). Marginal models for zero inflated clustered data. *Stat. Model.* **4** 161–180. MR2062098 https://doi.org/10.1191/1471082X04st076oa

HILBE, J. M. (2007). *Negative Binomial Regression*. Cambridge Univ. Press, Cambridge. MR2359854 https://doi.org/10.1017/CBO9780511811852

HINDE, J. and DEMÉTRIO, C. G. B. (1998). Overdispersion: Models and estimation. *Comput. Statist. Data Anal.* **27** 151–170.

HOLLA, M. S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika* **11** 115–121. MR0214180 https://doi.org/10.1007/BF02613581

KANG, T., LEVY, S. M. and DATTA, S. (2021). Analyzing longitudinal clustered count data with zero inflation: Marginal modeling using the Conway–Maxwell–Poisson distribution. *Biom. J.* **63** 761–786. MR4248729 https://doi.org/10.1002/bimj.202000061

KLEIBER, C. and ZEILEIS, A. (2008). *Applied Econometrics with R*. Springer, New York.

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

LAWLESS, J. F. (1987). Negative binomial and mixed Poisson regression. *Canad. J. Statist.* **15** 209–225. MR0926553 https://doi.org/10.2307/3314912

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

MA, R., HASAN, M. T. and SNEDDON, G. (2009). Modelling heterogeneity in clustered count data with extra zeros using compound Poisson random effect. *Stat. Med.* **28** 2356–2369. MR2751541 https://doi.org/10.1002/sim.3619

RIDOUT, M., HINDE, J. and DEMÉTRIO, C. G. B. (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometrics Conference. Cape Town, Invited Papers* 179–192.

RIDOUT, M., HINDE, J. and DEMÉTRIO, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57** 219–223. MR1833310 https://doi.org/10.1111/j.0006-341X.2001.00219.x

ROSEN, O., JIANG, W. and TANNER, M. A. (2000). Mixtures of marginal models. *Biometrika* **87** 391–404. MR1782486 https://doi.org/10.1093/biomet/87.2.391

SELLERS, K. F. and SHMUELI, G. (2010). A flexible regression model for count data. *Ann. Appl. Stat.* **4** 943–961. MR2758428 https://doi.org/10.1214/09-AOAS306

SHOUKRI, M. M., ASYALI, M. H., VANDORP, R. and KELTON, D. (2004). The Poisson inverse Gaussian regression model in the analysis of clustered counts data. *J. Data Sci.* **2** 17–32.

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

WILLMOT, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scand. Actuar. J.* **3–4** 113–127. MR0943576 https://doi.org/10.1080/03461238.1987.10413823

YAU, K. K. W., WANG, K. and LEE, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biom. J.* **45** 437–452. MR1984622 https://doi.org/10.1002/bimj.200390024

ZHELONKIN, M. and RONCHETTI, E. (2021). Robust analysis of sample selection models through the R package ssmrob. *J. Stat. Softw.* **99** 1–35.

ZHU, H., LEE, S.-Y., WEI, B.-C. and ZHOU, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika* **88** 727–737. MR1859405 https://doi.org/10.1093/biomet/88.3.727

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3364–3382
https://doi.org/10.1214/24-AOAS1940

MODELING URBAN CRIME OCCURRENCES VIA NETWORK REGULARIZED REGRESSION

BY ELIZABETH UPTON[1,a] AND LUIS CARVALHO[2,b]

[1]*Department of Mathematics and Statistics, Williams College,* [a]*emu1@williams.edu*
[2]*Department of Mathematics and Statistics, Boston University,* [b]*lecarval@bu.edu*

Analyses of occurrences of residential burglary in urban areas have shown that crime rates are not spatially homogeneous: rates vary across the network of city streets, resulting in some areas being far more susceptible to crime than others. The explanation for why a certain segment of the city experiences high crime may be different than why a neighboring area experiences high crime. Motivated by the importance of understanding spatial patterns such as these, we consider a statistical model of burglary defined on the street network of Boston, Massachusetts. Leveraging ideas from functional data analysis, our proposed solution consists of a generalized linear model with vertex-indexed covariates, allowing for an interpretation of the covariate effects at the street level. We employ a regularization procedure cast as a prior distribution on the regression coefficients under a Bayesian setup so that the predicted responses vary smoothly according to the connectivity of the city. We introduce a novel variable selection procedure, examine computationally efficient methods for sampling from the posterior distribution of the model parameters, and demonstrate the flexibility of our proposed modeling structure. The resulting model and interpretations provide insight into the spatial network patterns and dynamics of residential burglary in Boston.

REFERENCES

BALOCCHI, C. and JENSEN, S. T. (2019). Spatial modeling of trends in crime over time in Philadelphia. *Ann. Appl. Stat.* **13** 2235–2259. MR4037429 https://doi.org/10.1214/19-aoas1280

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

BELKIN, M., MATVEEVA, I. and NIYOGI, P. (2004). Regularization and semi-supervised learning on large graphs. In *Learning Theory. Lecture Notes in Computer Science* **3120** 624–638. Springer, Berlin. MR2177939 https://doi.org/10.1007/978-3-540-27819-1_43

BERNASCO, W. and BLOCK, R. (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology* **47** 93–130.

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. With discussion and a reply by Besag. MR1105822 https://doi.org/10.1007/BF00116466

BOWERS, K. J., JOHNSON, S. D. and PEASE, K. (2004). Prospective hot-spotting: The future of crime mapping? *Br. J. Criminol.* **44** 641–658.

CITY OF BOSTON (2016). Data Boston. Available at https://data.cityofboston.gov/. Retrieved February 16, 2016.

CITY OF BOSTON (2022). Boston Housing Conditions and Real Estate Trends Report. Available at https://www.bostonplans.org/getattachment/066b23c5-cab9-4731-a338-f6e57e3ef55f. Retrieved January 8, 2023.

CLEVELAND, C., STANTON, L., WOODS, B., MARTIN, A., FORTUNE, D., WALSH, M., CASTIGLIEGO, J., PEREZ, T., GALANTE, E. et al. (2019). Carbon Free Boston: Social equity report 2019.

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2001). *Introduction to Algorithms*, 2nd ed. MIT Press, Cambridge, MA, McGraw-Hill, Boston, MA. MR1848805

DAVIES, T. and JOHNSON, S. D. (2015). Examining the relationship between road structure and burglary risk via quantitative network analysis. *J. Quant. Criminol.* **31** 481–507.

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

Key words and phrases. Graph Laplacian, network inference, residential burglary.

DÖRFLER, F. and BULLO, F. (2013). Kron reduction of graphs with applications to electrical networks. *IEEE Trans. Circuits Syst. I. Regul. Pap.* **60** 150–163. MR3017573 https://doi.org/10.1109/TCSI.2012.2215780

ECK, J., CHAINEY, S., CAMERON, J. and WILSON, R. (2005). Mapping Crime: Understanding Hotspots Technical Report, National Institute of Justice.

FRITH, M. J., JOHNSON, S. D. and FRY, H. M. (2017). Role of the street network in Burglars' Spatial decision-making. *Criminology* **55** 344–376.

GARNER, B. A. (2001). *A Dictionary of Modern Legal Usage*. Oxford Univ. Press, USA.

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

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. With discussion and a reply by the authors. MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x

JOHNSON, S. D., GUERETTE, R. T. and BOWERS, K. (2014). Crime displacement: What we know, what we don't know, and what it means for crime reduction. *Journal of Experimental Criminology* **10** 549–571.

KIM, S., JOSHI, P., KALSI, P. S. and TAHERI, P. (2018). Crime analysis through machine learning. In 2018 *IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference* (*IEMCON*) 415–420. IEEE Press, New York.

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

LANCKRIET, G. R. G., BIE, T. D., CRISTIANINI, N., JORDAN, M. I. and NOBLE, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics* **20** 2626–2635. https://doi.org/10.1093/bioinformatics/bth294

LESKOVEC, J. and KREVL, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. Available at http://snap.stanford.edu/data. Retrieved March 23, 2017.

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

LI, T., LEVINA, E. and ZHU, J. (2019). Prediction models for network-linked data. *Ann. Appl. Stat.* **13** 132–164. MR3937424 https://doi.org/10.1214/18-AOAS1205

MAHFOUD, M., BHULAI, S., VAN DER MEI, R., ERKIN, D. and DUGUNDJI, E. (2019). Network analysis of city streets: Forecasting burglary risk in small areas. *Int. J. Adv. Secur.*

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. *Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition of [MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MEIJER, A. and WESSELS, M. (2019). Predictive policing: Review of benefits and drawbacks. *Int. J. Public Adm.* **42** 1031–1039.

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

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

OPEN DATA (2016). Boston Maps. Available at http://bostonopendata-boston.opendata.arcgis.com/. Retrieved February 16, 2016.

PALMER, C. J., PATHAK, P. A. et al. (2017). Gentrification and the amenity value of crime reductions: Evidence from rent deregulation Technical Report, National Bureau of Economic Research.

PORTA, S., CRUCITTI, P. and LATORA, V. (2006). The network analysis of urban streets: A dual approach. *Phys. A, Stat. Mech. Appl.* **369** 853–866.

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

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. MR1707311 https://doi.org/10.1007/978-1-4757-3071-5

ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753 https://doi.org/10.1080/01621459.2013.869223

SMOLA, A. J. and KONDOR, R. (2003). Kernels and regularization on graphs. In *Learning Theory and Kernel Machines* 144–158. Springer, Berlin.

UPTON, E. and CARVALHO, L. (2024). Supplement to "Modeling Urban Crime Occurrences via Network Regularized Regression." https://doi.org/10.1214/24-AOAS1940SUPPA, https://doi.org/10.1214/24-AOAS1940SUPPB

VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ 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

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3383–3404
https://doi.org/10.1214/24-AOAS1941

PREDICTING COVID-19 HOSPITALISATION USING A MIXTURE OF BAYESIAN PREDICTIVE SYNTHESES

BY GENYA KOBAYASHI[1,a], SHONOSUKE SUGASAWA[2,b], YUKI KAWAKUBO[3,c], DONGU HAN[4,d] AND TAERYON CHOI[4,e]

[1]*School of Commerce, Meiji University,* [a]*gkobayashi@meiji.ac.jp*
[2]*Faculty of Economics, Keio University,* [b]*sugasawa@econ.keio.ac.jp*
[3]*Graduate School of Social Sciences, Chiba University,* [c]*kawakubo@chiba-u.jp*
[4]*Department of Statstics, Korea University,* [d]*idontknow35@korea.ac.kr,* [e]*trchoi@korea.ac.kr*

This paper proposes a novel methodology called the mixture of Bayesian predictive syntheses (MBPS) for multiple time series count data for the challenging task of predicting the numbers of COVID-19 inpatients and isolated cases in Japan and Korea at the subnational level. MBPS combines a set of predictive models and partitions the multiple time series into clusters based on their contribution to predicting the outcome. In this way MBPS leverages the shared information within each cluster and is suitable for predicting COVID-19 inpatients since the data exhibit similar dynamics over multiple areas. Also, MBPS avoids using a multivariate count model, which is generally cumbersome to develop and implement. Our Japanese and Korean data analyses demonstrate that the proposed MBPS methodology has improved predictive accuracy and uncertainty quantification.

REFERENCES

AASTVEIT, K. A., MITCHELL, J., RAVAZZOLO, F. and VAN DIJK, H. K. (2019). The evolution of forecast density combinations in economics. In *Oxford Research Encyclopedia of Economics and Finance*.

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

CABEL, D., SUGASAWA, S., KATO, M., TAKANASHI, K. and MCALINN, K. (2023). Bayesian spatial predictive synthesis. arXiv preprint arXiv:2203.05197.

CHERNIS, T. (2024). Combining large numbers of density predictions with Bayesian predictive synthesis. *Stud. Nonlinear Dyn. Econom.* **28** 293–317. MR4746687

CHOWELL, G., DAHAL, S., TARIQ, A., ROOSA, K., HYMAN, J. M. and LUO, R. (2022). An ensemble n-sub-epidemic modeling framework for short-term forecasting epidemic trajectories: Application to the COVID-19 pandemic in the USA. *PLoS Comput. Biol.* **18** e1010602.

D'ANGELO, L. and RAVISHANKER, A. (2023). Efficient posterior sampling for Bayesian Poisson regression. *J. Comput. Graph. Statist.* **32** 916–926. MR4641469 https://doi.org/10.1080/10618600.2022.2123337

DAVIS, R. A., FOKIANOS, K., HOLAN, S. H., JOE, H., LIVSEY, J., LUND, R., PIPIRAS, V. and RAVISHANKER, N. (2021). Count time series: A methodological review. *J. Amer. Statist. Assoc.* **116** 1533–1547. MR4309291 https://doi.org/10.1080/01621459.2021.1904957

DAVIS, R. A., HOLAN, S. H., LUND, R. and RAVISHANKER, N., eds. (2016). *Handbook of Discrete-Valued Time Series. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3642975

FISHER, J. D., PUELZ, D. W. and CARVALHO, C. M. (2020). Monotonic effects of characteristics on returns. *Ann. Appl. Stat.* **14** 1622–1650. MR4194241 https://doi.org/10.1214/20-AOAS1351

FOKIANOS, K. (2024). Multivariate Count Time Series Modelling. *Econom. Stat.* **31** 100–116. MR4764408 https://doi.org/10.1016/j.ecosta.2021.11.006

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

FRÜHWIRTH-SCHNATTER, S. (2011). Panel data analysis: A survey on model-based clustering of time series. *Adv. Data Anal. Classif.* **5** 251–280. MR2860101 https://doi.org/10.1007/s11634-011-0100-0

Key words and phrases. Clustering, count data, dynamic factor model, finite mixture model, Markov chain Monte Carlo, Pólya-gamma augmentation, state space model.

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

HAMURA, Y., IRIE, K. and SUGASAWA, S. (2021). Robust Bayesian modeling of counts with zero inflation and outliers: Theoretical robustness and efficient computation. arXiv preprint arXiv:2106.10503v2.

JOHNSON, M. C. and WEST, M. (2023). Bayesian predictive synthesis with outcome-dependent pools. arXiv preprint arXiv:1803.01984.

KOBAYASHI, G., SUGASAWA, S., KAWAKUBO, Y., HAN, D. and CHOI, T. (2024). Supplement to "Predicting COVID-19 hospitalisation using a mixture of Bayesian predictive syntheses." https://doi.org/10.1214/24-AOAS1941SUPP

LIN, A., ZHANG, Y., HENG, J., ALLSOP, S. A., TYE, K. M., JACOB, P. E. and BA, D. (2019). Clustering time series with nonlinear dynamics: A Bayesian non-parametric and particle-based approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 2476–2484. PMLR.

MCALINN, K. (2021). Mixed-frequency Bayesian predictive synthesis for economic nowcasting. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **70** 1143–1163. MR4347707 https://doi.org/10.1111/rssc.12500

MCALINN, K., AASTVEIT, K. A., NAKAJIMA, J. and WEST, M. (2020). Multivariate Bayesian predictive synthesis in macroeconomic forecasting. *J. Amer. Statist. Assoc.* **115** 1092–1110. MR4143452 https://doi.org/10.1080/01621459.2019.1660171

MCALINN, K. and WEST, M. (2019). Dynamic Bayesian predictive synthesis in time series forecasting. *J. Econometrics* **210** 155–169. MR3944768 https://doi.org/10.1016/j.jeconom.2018.11.010

MCCARTHY, D. and JENSEN, S. T. (2016). Power-weighted densities for time series data. *Ann. Appl. Stat.* **10** 305–334. MR3480498 https://doi.org/10.1214/15-AOAS893

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

PAIREAU, J., ANDRONICO, A., HOZÉ, N., LAYAN, M., CRÉPEY, P., ROUMAGNAC, A., LAVIELLE, M., BOËLLE, P.-Y. and CAUCHEMEZ, S. (2022). An ensemble model based on early predictors to forecast COVID-19 health care demand in France. *Proc. Natl. Acad. Sci. USA* **119** e2103302119.

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

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

RAHIMI, I., CHEN, F. and GANDOMI, A. H. (2023). A review on COVID-19 forecasting models. *Neural Comput. Appl.* **35** 23671–23681.

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

SUGASAWA, S., TAKANASHI, K., MCALINN, K. and AIROLDI, E. M. (2023). Bayesian causal synthesis for meta-inference on heterogenous treatment effects. arXiv preprint arXiv:2304.07726.

TAKANASHI, K. and MCALINN, K. (2023). Equivariant online predictions of non-stationary time series. arXiv preprint arXiv:1911.08662.

TALLMAN, E. and WEST, M. (2024). Bayesian predictive decision synthesis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 340–363. MR4754087 https://doi.org/10.1093/jrsssb/qkad109

WEST, M. (2020). Bayesian forecasting of multivariate time series: Scalability, structure uncertainty and decisions. *Ann. Inst. Statist. Math.* **72** 1–31. MR4052647 https://doi.org/10.1007/s10463-019-00741-3

WEST, M. and CROSSE, J. (1992). Modelling probabilistic agent opinion. *J. Roy. Statist. Soc. Ser. B* **54** 285–299. MR1157726

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

WEST, M., HARRISON, P. J. and MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Assoc.* **80** 73–97. With discussion. MR0786598

ZHU, H.-T. and ZHANG, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 3–16. MR2035755 https://doi.org/10.1046/j.1369-7412.2003.05379.x

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3405–3424
https://doi.org/10.1214/24-AOAS1942

LEARNING BRAIN CONNECTIVITY IN SOCIAL COGNITION WITH DYNAMIC NETWORK REGRESSION

BY MAOYU ZHANG[1,a], BIAO CAI[2,c], WENLIN DAI[3,d], DEHAN KONG[4,e], HONGYU ZHAO[5,f] AND JINGFEI ZHANG[1,b]

[1]*Goizueta Business School, Emory University,* [a]*maoyu.zhang@emory.edu,* [b]*jingfei.zhang@emory.edu*
[2]*Department of Management Sciences, City University of Hong Kong,* [c]*biao.cai@cityu.edu.hk*
[3]*Institute of Statistics and Big Data, Renmin University of China,* [d]*wenlin.dai@ruc.edu.cn*
[4]*Department of Statistical Sciences, University of Toronto,* [e]*dehan.kong@utoronto.ca*
[5]*Department of Biostatistics, Yale University,* [f]*hongyu.zhao@yale.edu*

Dynamic networks have been increasingly used to characterize brain connectivity that varies during resting and task states. In such characterizations a connectivity network is typically measured at each time point for a subject over a common set of nodes representing brain regions, together with rich subject-level information. A common approach to analyzing such data is an edge-based method that models the connectivity between each pair of nodes separately. However, such approach may have limited performance when the noise level is high and the number of subjects is limited, as it does not take advantage of the inherent network structure. To better understand if and how the subject-level covariates affect the dynamic brain connectivity, we introduce a semiparametric dynamic network response regression that relates a dynamic brain connectivity network to a vector of subject-level covariates. A key advantage of our method is to exploit the structure of dynamic imaging coefficients in the form of high-order tensors. We develop an efficient estimation algorithm and evaluate the efficacy of our approach through simulation studies. Finally, we present our results on the analysis of a task-related study on social cognition in the Human Connectome Project, where we identify known sex-specific effects on brain connectivity that cannot be inferred using alternative methods.

REFERENCES

ADOLPHS, R. (2009). The social brain: Neural basis of social knowledge. *Annu. Rev. Psychol.* **60** 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

ALAHMADI, A. A. (2021). Investigating the sub-regions of the superior parietal cortex using functional magnetic resonance imaging connectivity. *Insights Imaging* **12** 1–12.

BARCH, D. M., BURGESS, G. C., HARMS, M. P., PETERSEN, S. E., SCHLAGGAR, B. L., CORBETTA, M., GLASSER, M. F., CURTISS, S., DIXIT, S. et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80** 169–189.

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

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

BI, X., QU, A. and SHEN, X. (2018). Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.* **46** 3308–3333. MR3852653 https://doi.org/10.1214/17-AOS1659

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

CAI, B., ZHANG, J. and SUN, W. W. (2021). Jointly modeling and clustering tensors in high dimensions. arXiv preprint. Available at arXiv:2104.07773.

CASTELLI, F., HAPPÉ, F., FRITH, U. and FRITH, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* **12** 314–325. https://doi.org/10.1006/nimg.2000.0612

Key words and phrases. Dynamic brain connectivity, low rank, network regression, tensors, theory of mind.

CHEN, J. and CHEN, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statist. Sinica* **22** 555–574. MR2954352 https://doi.org/10.5705/ss.2010.216

DENG, Y., TANG, X. and QU, A. (2023). Correlation tensor decomposition and its application in spatial imaging data. *J. Amer. Statist. Assoc.* **118** 440–456. MR4571133 https://doi.org/10.1080/01621459.2021.1938083

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

FISKE, S. T. and TAYLOR, S. E. (1991). *Social Cognition*. Mcgraw-Hill, New York.

GALLAGHER, H. L. and FRITH, C. D. (2003). Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* **7** 77–83. https://doi.org/10.1016/s1364-6613(02)00025-6

GOLDENBERG, G., PODREKA, I., STEINER, M., FRANZEN, P. and DEECKE, L. (1991). Contributions of occipital and temporal brain regions to visual and acoustic imagery–a spect study. *Neuropsychologia* **29** 695–702. https://doi.org/10.1016/0028-3932(91)90103-f

GOLDFARB, E. V., SEO, D. and SINHA, R. (2019). Sex differences in neural stress responses and correlation with subjective stress and stress regulation. *Neurobiol. Stress* **11** 100177. https://doi.org/10.1016/j.ynstr.2019.100177

HAO, B., WANG, B., WANG, P., ZHANG, J., YANG, J. and SUN, W. W. (2021). Sparse tensor additive regression. *J. Mach. Learn. Res.* **22** Paper No. 64, 43. MR4253757

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

HU, W., PAN, T., KONG, D. and SHEN, W. (2021). Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics* **77** 1227–1240. MR4357833 https://doi.org/10.1111/biom.13362

INGALHALIKAR, M., SMITH, A., PARKER, D., SATTERTHWAITE, T. D., ELLIOTT, M. A., RUPAREL, K., HAKONARSON, H., GUR, R. E., GUR, R. C. et al. (2014). Sex differences in the structural connectome of the human brain. *Proc. Natl. Acad. Sci. USA* **111** 823–828.

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

KONG, D., AN, B., ZHANG, J. and ZHU, H. (2020). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *J. Amer. Statist. Assoc.* **115** 403–424. MR4078472 https://doi.org/10.1080/01621459.2018.1555092

LIEBERMAN, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annu. Rev. Psychol.* **58** 259–289. https://doi.org/10.1146/annurev.psych.58.110405.085654

MATHER, M., LIGHTHALL, N. R., NGA, L. and GORLICK, M. A. (2010). Sex differences in how stress affects brain activity during face viewing. *NeuroReport* **21** 933–937. https://doi.org/10.1097/WNR.0b013e32833ddd92

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

PENSKY, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.* **47** 2378–2403. MR3953455 https://doi.org/10.1214/18-AOS1751

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

PROVERBIO, A. M., ADORNI, R., ZANI, A. and TRESTIANU, L. (2009). Sex differences in the brain response to affective scenes with or without humans. *Neuropsychologia* **47** 2374–2388. https://doi.org/10.1016/j.neuropsychologia.2008.10.030

SAXE, R. and KANWISHER, N. (2013). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". In *Social Neuroscience* 171–182. Psychology Press, London.

SCHNABEL, R. B., KOONATZ, J. E. and WEISS, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Trans. Math. Software* **11** 419–440. MR0828567 https://doi.org/10.1145/6187.6192

SCHÖLVINCK, M. L., MAIER, A., YE, F. Q., DUYN, J. H. and LEOPOLD, D. A. (2010). Neural basis of global resting-state fMRI activity. *Proc. Natl. Acad. Sci. USA* **107** 10238–10243.

SMITH, S. M., VIDAURRE, D., BECKMANN, C. F., GLASSER, M. F., JENKINSON, M., MILLER, K. L., NICHOLS, T. E., ROBINSON, E. C., SALIMI-KHORSHIDI, G. et al. (2013). Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17** 666–682.

SRIVASTAVA, S., ENGELHARDT, B. E. and DUNSON, D. B. (2017). Expandable factor analysis. *Biometrika* **104** 649–663. MR3694588 https://doi.org/10.1093/biomet/asx030

SUN, W. W. and LI, L. (2017). STORE: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** Paper No. 135, 37. MR3763769

SUN, W. W., LU, J., LIU, H. and CHENG, G. (2017). Provable sparse tensor decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 899–916. MR3641413 https://doi.org/10.1111/rssb.12190

TANG, X., BI, X. and QU, A. (2020). Individualized multilayer tensor learning with an application in imaging analysis. *J. Amer. Statist. Assoc.* **115** 836–851. MR4107683 https://doi.org/10.1080/01621459.2019.1585254

WANG, J., KORCZYKOWSKI, M., RAO, H., FAN, Y., PLUTA, J., GUR, R. C., MCEWEN, B. S. and DETRE, J. A. (2007). Gender difference in neural response to psychological stress. *Soc. Cogn. Affect. Neurosci.* **2** 227–239. https://doi.org/10.1093/scan/nsm018

WANG, L., DURANTE, D., JUNG, R. E. and DUNSON, D. B. (2017). Bayesian network-response regression. *Bioinformatics* **33** 1859–1866. https://doi.org/10.1093/bioinformatics/btx050

WHEATLEY, T., MILLEVILLE, S. C. and MARTIN, A. (2007). Understanding animate agents: Distinct roles for the social network and mirror system. *Psychol. Sci.* **18** 469–474. https://doi.org/10.1111/j.1467-9280.2007.01923.x

XU, K. S. and HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562.

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

ZHANG, J. and CAO, J. (2017). Finding common modules in a time-varying network with application to the *Drosophila melanogaster* gene regulation network. *J. Amer. Statist. Assoc.* **112** 994–1008. MR3735355 https://doi.org/10.1080/01621459.2016.1260465

ZHANG, J., SUN, W. W. and LI, L. (2020). Mixed-effect time-varying network model and application in brain connectivity analysis. *J. Amer. Statist. Assoc.* **115** 2022–2036. MR4189774 https://doi.org/10.1080/01621459.2019.1677242

ZHANG, J., SUN, W. W. and LI, L. (2023). Generalized connectivity matrix response regression with applications in brain connectivity studies. *J. Comput. Graph. Statist.* **32** 252–262. MR4552951 https://doi.org/10.1080/10618600.2022.2074434

ZHANG, M., CAI, B., DAI, W., KONG, D., ZHAO, H. and ZHANG, J. (2024). Supplement to "Learning Brain Connectivity in Social Cognition with Dynamic Network Regression." https://doi.org/10.1214/24-AOAS1942SUPPA, https://doi.org/10.1214/24-AOAS1942SUPPB

ZHANG, X. and LI, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59** 426–436. MR3740960 https://doi.org/10.1080/00401706.2016.1272495

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

ZHOU, J., SUN, W. W., ZHANG, J. and LI, L. (2021). Partially observed dynamic tensor response regression. *J. Amer. Statist. Assoc.* **118** 424–439. MR4571132 https://doi.org/10.1080/01621459.2021.1938082

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3425–3443
https://doi.org/10.1214/24-AOAS1943

MODELING TRAJECTORIES USING FUNCTIONAL LINEAR DIFFERENTIAL EQUATIONS

BY JULIA WROBEL[1,a], BRITTON SAUERBREI[2,b], ERIC A. KIRK[2,c], JIAN-ZHONG GUO[3,d], ADAM HANTMAN[3,e] AND JEFF GOLDSMITH[4,f]

[1]*Department of Biostatistics and Bioinformatics, Emory University,* [a]*julia.wrobel@emory.edu*
[2]*Department of Neurosciences, Case Western Reserve University,* [b]*bxs561@case.edu,* [c]*eak152@case.edu*
[3]*Department of Cell Biology and Physiology, UNC Medical Center,* [d]*jayzhong@email.unc.edu,* [e]*adam_hantman@med.unc.edu*
[4]*Department of Biostatistics, Columbia University Mailman School of Public Health,* [f]*jeff.goldsmith@columbia.edu*

We are motivated by a study that seeks to better understand the dynamic relationship between muscle activation and paw position during locomotion. For each gait cycle in this experiment, activation in the biceps and triceps is measured continuously and in parallel with paw position as a mouse trotted on a treadmill. We propose an innovative general regression method that draws from both ordinary differential equations and functional data analysis to model the relationship between these functional inputs and responses as a dynamical system that evolves over time. Specifically, our model addresses gaps in both literatures and borrows strength across curves estimating ODE parameters across all curves simultaneously rather than separately modeling each functional observation. Our approach compares favorably to related functional data methods in simulations and in cross-validated predictive accuracy of paw position in the gait data. In the analysis of the gait cycles, we find that paw speed and position are dynamically influenced by inputs from the biceps and triceps muscles and that the effect of muscle activation persists beyond the activation itself.

REFERENCES

ADEYEFA, E. O. (2021). A model for solving first, second and third order IVPs directly. *Int. J. Appl. Comput. Math.* **7** 131.

BECKER, M. I., CALAME, D. J., WROBEL, J. and PERSON, A. L. (2020). Online control of reach accuracy in mice. *J. Neurophysiol.* **124** 1637–1655. https://doi.org/10.1152/jn.00324.2020

BEKINS, B. A., WARREN, E. and GODSY, E. M. (1998). A comparison of zero-order, first-order, and Monod biotransformation models. *Groundwater* **36** 261–268.

BORCHERS, H. W. and BORCHERS, M. H. W. (2021). Package 'pracma'.

CHEN, S., SHOJAIE, A. and WITTEN, D. M. (2017). Network reconstruction from high-dimensional ordinary differential equations. *J. Amer. Statist. Assoc.* **112** 1697–1707.

DAI, X. and LI, L. (2022). Kernel ordinary differential equations. *J. Amer. Statist. Assoc.* **117** 1711–1725.

DAI, X. and LI, L. (2024). Post-regularization confidence bands for ordinary differential equations. *J. Mach. Learn. Res.* **25** Paper No. [23], 51. MR4723873

DATTNER, I. and KLAASSEN, C. A. J. (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electron. J. Stat.* **9** 1939–1973. MR3391125 https://doi.org/10.1214/15-EJS1053

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121.

FAN, J. and ZHANG, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface* **1** 179.

GOLDSMITH, J. (2016). vbvs.concurrent: Fitting methods for the functional linear concurrent model. *J. Open Sour. Softw.* **1** 141.

GOLDSMITH, J. and KITAGO, T. (2016). Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 215–236.

Key words and phrases. Functional regression, ordinary differential equations, nonlinear least squares, dynamical systems.

GOLDSMITH, J., SCHEIPL, F., HUANG, L., WROBEL, J., GELLAR, J., HAREZLAK, J., MCLEAN, M. W., SWIHART, B., XIAO, L. et al. (2021). Refund: regression with functional data. R package version 0.1-24.

GOLDSMITH, J. and SCHWARTZ, J. E. (2017). Variable selection in the functional linear concurrent model. *Stat. Med.* **36** 2237–2250. https://doi.org/10.1002/sim.7254

GUNNING, E. and HOOKER, G. (2024). An understanding of principal differential analysis. arXiv preprint. Available at arXiv:2406.18484.

GUO, J.-Z., GRAVES, A. R., GUO, W. W., ZHENG, J., LEE, A., RODRIGUEZ-GONZALEZ, J., LI, N., MACKLIN, J. J., PHILLIPS, J. W. et al. (2015). Cortex commands the performance of skilled movement. *eLife* **4** e10774.

HENDERSON, J. and MICHAILIDIS, G. (2014). Network reconstruction using nonparametric additive ODE models. *PLoS ONE* **9** e94003. https://doi.org/10.1371/journal.pone.0094003

KIRK, E. A., HOPE, K. T., SOBER, S. J. and SAUERBREI, B. A. (2023). An output-null signature of inertial load in motor cortex. bioRxiv 2023–11.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LEROUX, A., XIAO, L., CRAINICEANU, C. and CHECKLEY, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. *Stat. Med.* **37** 1376–1388. https://doi.org/10.1002/sim.7582

LOVELAND, W. D. MORRISSEY, D. J. and SEABORG, G. T. (2017). *Modern Nuclear Chemistry*. Wiley, New York.

LU, T., LIANG, H., LI, H. and WU, H. (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *J. Amer. Statist. Assoc.* **106** 1242–1258.

MALFAIT, N. and RAMSAY, J. O. (2003). The historical functional linear model. *Canad. J. Statist.* **31** 115–128.

NELSON, P. W. and PERELSON, A. S. (2002). Mathematical analysis of delay differential equation models of HIV-1 infection. *Math. Biosci.* **179** 73–94. https://doi.org/10.1016/s0025-5564(02)00099-8

PISTOHL, T., BALL, T., SCHULZE-BONHAGE, A., AERTSEN, A. and MEHRING, C. (2008). Prediction of arm movement trajectories from ECoG-recordings in humans. *J. Neurosci. Methods* **167** 105–114.

RAMSAY, J. and HOOKER, G. (2017). *Dynamic Data Analysis*. Springer, Berlin.

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.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer, New York.

RAO, A. R. and REIMHERR, M. (2023). Modern non-linear function-on-function regression. *Stat. Comput.* **33** Paper No. 130, 12. MR4650055 https://doi.org/10.1007/s11222-023-10299-z

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge.

SAUERBREI, B. A., GUO, J.-Z., COHEN, J. D., MISCHIATI, M., GUO, W., KABRA, M., VERMA, N., MENSH, B., BRANSON, K. and HANTMAN, A. W. (2020). Cortical pattern generation during dexterous movement is input-driven. *Nature* **577** 386–391.

SCHEIPL, F., GERTHEISS, J. and GREVEN, S. (2016). Generalized functional additive mixed models. *Electron. J. Stat.* **10** 1455–1492. MR3507370 https://doi.org/10.1214/16-EJS1145

SCHEIPL, F., STAICU, A.-M. and GREVEN, S. (2015). Functional additive mixed models. *J. Comput. Graph. Statist.* **24** 477–501.

TENNENBAUM, M. and POLLARD, H. (1985). *Ordinary Differential Equations*: *An Elementary Textbook for Students of Mathematics*, *Engineering*, *and the Sciences*.

WALKER, S. (1996). An EM algorithm for nonlinear random effects models. *Biometrics* **52** 934–944. MR1411741 https://doi.org/10.2307/2533054

WROBEL, J., SAUERBREI, B., KIRK, E. A., GUO, J.-Z., HANTMAN, A. and GOLDSMITH, J. (2024). Supplement to "Modeling trajectories using functional linear differential equations." https://doi.org/10.1214/24-AOAS1943SUPPA, https://doi.org/10.1214/24-AOAS1943SUPPB

WU, W. and SRIVASTAVA, A. (2014). Analysis of spike train data: Alignment and comparisons using the extended Fisher–Rao metric. *Electron. J. Stat.* **8** 1776–1785.

YAO, F., MÜLLER, H. G. and WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590.

YAVUZTURK, C., SPITLER, J. D., REES, S. J. et al. (1999). A transient two-dimensional finite volume model for the simulation of vertical U-tube ground heat exchangers. *ASHRAE Trans.* **105** 465–474.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3444–3466
https://doi.org/10.1214/24-AOAS1944

A SPATIALLY VARYING HIERARCHICAL RANDOM EFFECTS MODEL FOR LONGITUDINAL MACULAR STRUCTURAL DATA IN GLAUCOMA PATIENTS

BY ERICA SU[1,a], ROBERT E. WEISS[1,b], KOUROS NOURI-MAHDAVI[2,d] AND ANDREW J. HOLBROOK[1,c]

[1]*Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles,* [a]*ericasu@ucla.edu,* [b]*robweiss@ucla.edu,* [c]*aholbroo@g.ucla.edu*

[2]*Glaucoma Division, Stein Eye Institute, David Geffen School of Medicine, University of California, Los Angeles,* [d]*nouri-mahdavi@jsei.ucla.edu*

We model longitudinal macular thickness measurements to monitor the course of glaucoma and prevent vision loss due to disease progression. The macular thickness varies over a 6×6 grid of locations on the retina, with additional variability arising from the imaging process at each visit. Currently, ophthalmologists estimate slopes using repeated simple linear regression for each subject and location. To estimate slopes more precisely, we develop a novel Bayesian hierarchical model for multiple subjects with spatially varying population-level and subject-level coefficients, borrowing information over subjects and measurement locations. We augment the model with visit effects to account for observed spatially correlated visit-specific errors. We model spatially varying: (a) intercepts, (b) slopes, and (c) log-residual standard deviations (SD) with multivariate Gaussian process priors with Matérn cross-covariance functions. Each marginal process assumes an exponential kernel with its own SD and spatial correlation matrix. We develop our models for and apply them to data from the Advanced Glaucoma Progression Study. We show that including visit effects in the model reduces error in predicting future thickness measurements and greatly improves model fit.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. *National Bureau of Standards Applied Mathematics Series*, *No*. 55. U. S. Government Printing Office, Washington, DC. MR0167642

APANASOVICH, T. V., GENTON, M. G. and SUN, Y. (2012). A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *J. Amer. Statist. Assoc*. **107** 180–193. MR2949350 https://doi.org/10.1080/01621459.2011.643197

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

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

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

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

BETZ-STABLEIN, B. D., MORGAN, W. H., HOUSE, P. H. and HAZELTON, M. L. (2013). Spatial modeling of visual field data for assessing glaucoma progression. *Investig. Ophthalmol. Vis. Sci*. **54** 1544–1553.

BOGACHEV, V. I. (1998). *Gaussian Measures*. *Mathematical Surveys and Monographs* **62**. Amer. Math. Soc., Providence, RI. MR1642391 https://doi.org/10.1090/surv/062

Key words and phrases. Bayesian modeling, ganglion cell complex, glaucoma, multivariate Gaussian processes, optical coherence tomography, random effects, spatially varying coefficients.

BRYAN, S. R., EILERS, P. H., LESAFFRE, E. M., LEMIJ, H. G. and VERMEER, K. A. (2015). Global visit effects in point-wise longitudinal modeling of glaucomatous visual fields. *Investig. Ophthalmol. Vis. Sci.* **56** 4283–4289.

BRYAN, S. R., EILERS, P. H. C., VAN ROSMALEN, J., RIZOPOULOS, D., VERMEER, K. A., LEMIJ, H. G. and LESAFFRE, E. M. E. H. (2017). Bayesian hierarchical modeling of longitudinal glaucomatous visual fields using a two-stage approach. *Stat. Med.* **36** 1735–1753. MR3648619 https://doi.org/10.1002/sim.7235

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

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

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

GARDINER, S. K. and CRABB, D. P. (2002). Examination of different pointwise linear regression methods for determining visual field progression. *Investig. Ophthalmol. Vis. Sci.* **43** 1400–1407.

GASPARI, G. and COHN, S. E. (1999). Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **125** 723–757.

GE, T., MÜLLER-LENKE, N., BENDFELDT, K., NICHOLS, T. E. and JOHNSON, T. D. (2014). Analysis of multiple sclerosis lesions via spatially varying coefficients. *Ann. Appl. Stat.* **8** 1095–1118. MR3262547 https://doi.org/10.1214/14-AOAS718

GELFAND, A. E., DIGGLE, P. J., FUENTES, M. and GUTTORP, P., eds. (2010) *Handbook of Spatial Statistics. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR2761512 https://doi.org/10.1201/9781420072884

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

GELFAND, A. E. and SCHLIEP, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spat. Stat.* **18** 86–104. MR3573271 https://doi.org/10.1016/j.spasta.2016.03.006

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

GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. MR3353096 https://doi.org/10.1214/14-STS487

GHITA, A. M., ILIESCU, D. A., GHITA, A. C., ILIE, L. A. and OTOBIC, A. (2023). Ganglion cell complex analysis: Correlations with retinal nerve fiber layer on optical coherence tomography. *Diagnostics* **13** 266.

GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. MR2752612 https://doi.org/10.1198/jasa.2010.tm09420

GÖSSL, C., AUER, D. P. and FAHRMEIR, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics* **57** 554–562. MR1855691 https://doi.org/10.1111/j.0006-341X.2001.00554.x

GUTTORP, P. and GNEITING, T. (2006). Studies in the history of probability and statistics XLIX On the Matérn correlation family. *Biometrika* **93** 989–995. MR2285084 https://doi.org/10.1093/biomet/93.4.989

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HOLLÓ, G. and NAGHIZADEH, F. (2015). Influence of a new software version of the RTVue-100 optical coherence tomograph on the detection of glaucomatous structural progression. *Eur. J. Ophthalmol.* **25** 410–415. https://doi.org/10.5301/ejo.5000576

JIN, X., CARLIN, B. P. and BANERJEE, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61** 950–961. MR2216188 https://doi.org/10.1111/j.1541-0420.2005.00359.x

KIM, H. and LEE, J. (2017). Hierarchical spatially varying coefficient process model. *Technometrics* **59** 521–527. MR3740968 https://doi.org/10.1080/00401706.2017.1317290

KINGMAN, S. (2004). Glaucoma is second leading cause of blindness globally. *Bull. World Health Organ.* **82** 887–888.

LEUNG, C. K., YE, C., WEINREB, R. N., YU, M., LAI, G. and LAM, D. S. (2013). Impact of age-related change of retinal nerve fiber layer and macular thicknesses on evaluation of glaucoma progression. *Ophthalmology* **120** 2485–2492.

LIU, Z., BARTSCH, A. J., BERROCAL, V. J. and JOHNSON, T. D. (2019). A mixed-effects, spatially varying coefficients model with application to multi-resolution functional magnetic resonance imaging data. *Stat. Methods Med. Res.* **28** 1203–1215. MR3934644 https://doi.org/10.1177/0962280217752378

MacNab, Y. C. (2016a). Linear models of coregionalization for multivariate lattice data: A general framework for coregionalized multivariate CAR models. *Stat. Med.* **35** 3827–3850. MR3538050 https://doi.org/10.1002/sim.6955

MacNab, Y. C. (2016b). Linear models of coregionalization for multivariate lattice data: Order-dependent and order-free cMCARs. *Stat. Methods Med. Res.* **25** 1118–1144. MR3541088 https://doi.org/10.1177/0962280216660419

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

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

Miraftabi, A., Amini, N., Gornbein, J., Henry, S., Romero, P., Coleman, A. L., Caprioli, J. and Nouri-Mahdavi, K. (2016). Local variability of macular thickness measurements with SD-OCT and influencing factors. *Transl. Vis. Sci. Technol.* **5** 5.

Mohammadzadeh, V., Fatehi, N., Yarmohammadi, A., Lee, J. W., Sharifipour, F., Daneshvar, R., Caprioli, J. and Nouri-Mahdavi, K. (2020a). Macular imaging with optical coherence tomography in glaucoma. *Surv. Ophthalmol.* **65** 597–638.

Mohammadzadeh, V., Rabiolo, A., Fu, Q., Morales, E., Coleman, A. L., Law, S. K., Caprioli, J. and Nouri-Mahdavi, K. (2020b). Longitudinal macular structure–function relationships in glaucoma. *Ophthalmology* **127** 888–900.

Mohammadzadeh, V., Su, E., Rabiolo, A., Shi, L., Zadeh, S. H., Law, S. K., Coleman, A. L., Caprioli, J., Weiss, R. E. et al. (2022a). Ganglion cell complex: The optimal measure for detection of structural progression in the macula. *Am. J. Ophthalmol.* **237** 71–82.

Mohammadzadeh, V., Su, E., Shi, L., Coleman, A. L., Law, S. K., Caprioli, J., Weiss, R. E. and Nouri-Mahdavi, K. (2022b). Multivariate longitudinal modeling of macular ganglion cell complex: Spatiotemporal correlations and patterns of longitudinal change. *Ophthalmol. Sci.* **2** 100187.

Mohammadzadeh, V., Su, E., Zadeh, S. H., Law, S. K., Coleman, A. L., Caprioli, J., Weiss, R. E. and Nouri-Mahdavi, K. (2021). Estimating ganglion cell complex rates of change with Bayesian hierarchical models. *Transl. Vis. Sci. Technol.* **10** 15.

Montesano, G., Garway-Heath, D. F., Ometto, G. and Crabb, D. P. (2021). Hierarchical censored Bayesian analysis of visual field progression. *Transl. Vis. Sci. Technol.* **10** 4.

Nishida, T., Moghimi, S., Mohammadzadeh, V., Wu, J.-H., Yamane, M. L., Kamalipour, A., Mahmoudinezhad, G., Micheletti, E., Liebmann, J. M. et al. (2022). Association between ganglion cell complex thinning and vision-related quality of life in glaucoma. *JAMA Ophthalmol.* **140** 800–806.

Nouri-Mahdavi, K., Hoffman, D., Ralli, M. and Caprioli, J. (2007). Comparison of methods to predict visual field progression in glaucoma. *Arch. Ophthalmol.* **125** 1176–1181. https://doi.org/10.1001/archopht.125.9.1176

Penny, W. D., Trujillo-Barreto, N. J. and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362. https://doi.org/10.1016/j.neuroimage.2004.08.034

Rabiolo, A., Mohammadzadeh, V., Fatehi, N., Morales, E., Coleman, A. L., Law, S. K., Caprioli, J. and Nouri-Mahdavi, K. (2020). Comparison of rates of progression of macular OCT measures in glaucoma. *Transl. Vis. Sci. Technol.* **9** 50.

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

Risser, M. D. and Turek, D. (2020). Bayesian inference for high-dimensional nonstationary Gaussian processes. *J. Stat. Comput. Simul.* **90** 2902–2928. MR4168232 https://doi.org/10.1080/00949655.2020.1792472

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2080278 https://doi.org/10.1007/978-1-4757-4145-2

Schmidt, A. M. and Gelfand, A. E. (2003). A Bayesian coregionalization approach for multivariate pollutant data. *J. Geophys. Res., Atmos.* **108**.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47. MR0501454

Su, E., Weiss, R. E, Nouri-Mahdavi, K. and Holbrook, A. J (2024). Supplement to "A spatially varying hierarchical random effects model for longitudinal macular structural data in glaucoma patients." https://doi.org/10.1214/24-AOAS1944SUPPA, https://doi.org/10.1214/24-AOAS1944SUPPB

Tan, O., Li, G., Lu, A. T.-H., Varma, R., Huang, D. and Advanced Imaging for Glaucoma Study Group (2008). Mapping of macular substructures with optical coherence tomography for glaucoma diagnosis. *Ophthalmology* **115** 949–956.

Tatham, A. J. and Medeiros, F. A. (2017). Detecting structural progression in glaucoma with optical coherence tomography. *Ophthalmology* **124** S57–S65. https://doi.org/10.1016/j.ophtha.2017.07.015

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

THOMPSON, A. C., JAMMAL, A. A., BERCHUCK, S. I., MARIOTTONI, E. B., WU, Z., DAGA, F. B., OGATA, N. G., URATA, C. N., ESTRELA, T. et al. (2020). Comparing the rule of 5 to trend-based analysis for detecting glaucoma progression on OCT. *Ophthalmol. Glaucoma* **3** 414–420.

TIBBITS, M. M., GROENDYKE, C., HARAN, M. and LIECHTY, J. C. (2014). Automated factor slice sampling. *J. Comput. Graph. Statist.* **23** 543–563. MR3215824 https://doi.org/10.1080/10618600.2013.791193

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27** 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

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

VER HOEF, J. M. and BARRY, R. P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction. *J. Statist. Plann. Inference* **69** 275–294. MR1631328 https://doi.org/10.1016/S0378-3758(97)00162-6

WACKERNAGEL, H. (2013). *Multivariate Geostatistics*, 3rd ed. Springer, Berlin.

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

WEINREB, R. N. and KHAW, P. T. (2004). Primary open-angle glaucoma. *Lancet* **363** 1711–1720. https://doi.org/10.1016/S0140-6736(04)16257-0

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

ZHANG, F., JIANG, W., WONG, P. and WANG, J.-P. (2016a). A Bayesian probit model with spatially varying coefficients for brain decoding using fMRI data. *Stat. Med.* **35** 4380–4397. MR3554969 https://doi.org/10.1002/sim.6999

ZHANG, X., FRANCIS, B. A., DASTIRIDOU, A., CHOPRA, V., TAN, O., VARMA, R., GREENFIELD, D. S., SCHUMAN, J. S., HUANG, D. et al. (2016b). Longitudinal and cross-sectional analyses of age effects on retinal nerve fiber layer and ganglion cell complex thickness by Fourier-domain OCT. *Transl. Vis. Sci. Technol.* **5** 1–1.

ZHU, H., FAN, J. and KONG, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Amer. Statist. Assoc.* **109** 1084–1098. MR3265682 https://doi.org/10.1080/01621459.2014.881742

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3467–3485
https://doi.org/10.1214/24-AOAS1945

MULTIPLE LATENT CLUSTERING MODEL FOR THE INFERENCE OF RNA LIFE-CYCLE KINETIC RATES FROM SEQUENCING DATA

BY GIANLUCA MASTRANTONIO[1,a], ENRICO BIBBONA[1,b] AND MATTIA FURLAN[2,c]

[1]*Department of Mathematical Sciences, Politecnico di Torino,* [a]*gianluca.mastrantonio@polito.it,* [b]*enrico.bibbona@polito.it*
[2]*Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milan, Italy,* [c]*mattia.furlan@iit.it*

We propose a hierarchical Bayesian model to infer RNA synthesis, processing, and degradation rates from time-course RNA sequencing data, based on an ordinary differential equation system that models the RNA life cycle. We parametrize the latent kinetic rates, which rule the system, with a novel functional form and estimate their parameters through three Dirichlet process mixture models. Owing to the complexity of this approach, we are able to simultaneously perform inference, clustering, and model selection. We apply our method to investigate transcriptional and post-transcriptional responses of murine fibroblasts to the activation of the proto-oncogene Myc. Our approach uncovers simultaneous regulations of the rates, which had been largely missed in previous analyses of this biological system.

REFERENCES

ALKALLAS, R., FISH, L., GOODARZI, H. and NAJAFABADI, H. S. (2017). Inference of rna decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease. *Nat. Commun.* **8**.

ALLOCCO, D. J., KOHANE, I. S. and BUTTE, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinform.* **5** 18.

ANDRIEU, C. and THOMS, J. (2008). A tutorial on adaptive MCMC. *Stat. Comput.* **18** 343–373. MR2461882 https://doi.org/10.1007/s11222-008-9110-y

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.

BANDIERA, R., WAGNER, R. E., BRITTO-BORGES, T., DIETERICH, C., DIETMANN, S., BORNELÖV, S. and FRYE, M. (2021). Rn7sk small nuclear rna controls bidirectional transcription of highly expressed gene pairs in skin. *Nat. Commun.* **12**.

BERGEN, V., LANGE, M., PEIDLI, S., WOLF, F. A. and THEIS, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38** 1408–1414.

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

CHECHIK, G. and KOLLER, D. (2009). Timing of gene expression responses to environmental changes. *J. Comput. Biol.* **16** 279–290.

CHEN, H., LIU, H. and QING, G. (2018). Targeting oncogenic myc as a strategy for cancer treatment. *Signal Transduct. Targeted Ther.* **3** 5.

CHEN, T. and VAN STEENSEL, B. (2017). Comprehensive analysis of nucleocytoplasmic dynamics of mrna in drosophila cells. *PLoS Genet.* **13** e1006929.

CHOI, J., LYSAKOVSKAIA, K., STIK, G., DEMEL, C., SÖDING, J., TIAN, T. V., GRAF, T. and CRAMER, P. (2021). Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *eLife* **10**. https://doi.org/10.7554/eLife.65381

CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M., GAFFNEY, D. J., ELO, L. L. et al. (2016). A survey of best practices for rna-seq data analysis. *Genome Biol.* **17** 13.

DANG, C. V. (2012). Myc on the path to cancer. *Cell* **149** 22–35.

DAVARI, K., LICHTI, J., GALLUS, C., GREULICH, F., UHLENHAUT, N. H., HEINIG, M., FRIEDEL, C. C. and GLASMACHER, E. (2017). Rapid genome-wide recruitment of rna polymerase ii drives transcription, splicing, and translation events during t cell responses. *Cell Rep.* **19** 643–654.

Key words and phrases. Dirichlet process, kinetic rates, RNA, Gene expression.

DE PRETIS, S., KRESS, T., MORELLI, M. J., MELLONI, G. E. M., RIVA, L., AMATI, B. and PELIZZOLA, M. (2015). INSPEcT: A computational tool to infer mRNA synthesis, processing and degradation dynamics from RNA- and 4sU-seq time course experiments. *Bioinformatics* **31** 2829–2835.

DE PRETIS, S., KRESS, T. R., MORELLI, M. J., SABÒ, A., LOCARNO, C., VERRECCHIA, A., DONI, M., CAMPANER, S., AMATI, B. et al. (2017). Integrative analysis of RNA polymerase II and transcriptional dynamics upon MYC activation. *Genome Res*. **27** 1658–1664. https://doi.org/10.1101/gr.226035.117

DESSIMOZ, C. and ŠKUNCA, N., eds. (2017). *The Gene Ontology Handbook*. *Methods in Molecular Biology* **1446**. Humana Press, New York. OCLC: ocn959227666.

DÖLKEN, L., RUZSICS, Z., RÄDLE, B., FRIEDEL, C. C., ZIMMER, R., MAGES, J., HOFFMANN, R., DICKINSON, P., FORSTER, T. et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of rna synthesis and decay. *RNA* **14** 1959–1972.

DUDEK, A. (2020). Silhouette index as clustering evaluation tool. In *Classification and Data Analysis* (K. Jajuga, J. Batóg and M. Walesiak, eds.) 19–33. Springer, Cham.

EDGAR, R., DOMRACHEV, M. and LASH, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. **30** 207–210.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc*. **90** 577–588. MR1340510

FANG, H., HUANG, Y.-F., RADHAKRISHNAN, A., SIEPEL, A., LYON, G. J. and SCHATZ, M. C. (2018). Scikit-ribo enables accurate estimation and robust modeling of translation dynamics at codon resolution. *Cell Syst*. **6** 180–191.e4.

FARINA, L., DE SANTIS, A., SALVUCCI, S., MORELLI, G. and RUBERTI, I. (2008). Embedding mrna stability in correlation analysis of time-series gene expression data. *PLoS Comput. Biol*. **4** 1–12.

FURLAN, M., GALEOTA, E., GAUDIO, N. D., DASSI, E., CASELLE, M., DE PRETIS, S. and PELIZZOLA, M. (2020). Genome-wide dynamics of RNA synthesis, processing, and degradation without RNA metabolic labeling. *Genome Res*. **30** 1492–1507. https://doi.org/10.1101/gr.260984.120

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC, Boca Raton, FL.

GNEDIN, A. and KEROV, S. (2001). A characterization of GEM distributions. *Combin. Probab. Comput*. **10** 213–217. MR1841641 https://doi.org/10.1017/S0963548301004692

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

GOODWIN, S., MCPHERSON, J. D. and MCCOMBIE, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet*. **17** 333–351.

HSU, T. Y.-T., SIMON, L. M., NEILL, N. J., MARCOTTE, R., SAYAD, A., BLAND, C. S., ECHEVERRIA, G. V., SUN, T., KURLEY, S. J. et al. (2015). The spliceosome is a therapeutic vulnerability in myc-driven cancer. *Nature* **525** 384–388.

HUANG, Y. and SANGUINETTI, G. (2016). Statistical modeling of isoform splicing dynamics from RNA-seq time series data. *Bioinformatics* **32** 2965–2972.

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

JAIN, S. and NEAL, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Anal*. **2** 445–472. MR2342168 https://doi.org/10.1214/07-BA219

JÜRGES, C., DÖLKEN, L. and ERHARD, F. (2018). Dissecting newly transcribed and old rna using grand-slam. *Bioinformatics* **34** i218–i226.

LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P. et al. (2018). Rna velocity of single cells. *Nature* **560** 494–498.

LI, G.-W. (2015). How do bacteria tune translation efficiency? *Curr. Opin. Microbiol*. **24** 66–71. https://doi.org/10.1016/j.mib.2015.01.001

LI, H.-B., TONG, J., ZHU, S., BATISTA, P. J., DUFFY, E. E., ZHAO, J., BAILIS, W., CAO, G., KROEHLING, L. et al. (2017). m6a mrna methylation controls t cell homeostasis by targeting the il-7/stat5/socs pathways. *Nature* **548** 338–342.

LITTLEWOOD, T. D., HANCOCK, D. C., DANIELIAN, P. S., PARKER, M. G. and EVAN, G. I. (1995). A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. *Nucleic Acids Res*. **23** 1686–1690.

LIU, H., ARSIÈ, R., SCHWABE, D., SCHILLING, M., MINIA, I., ALLES, J., BOLTENGAGEN, A., KOCKS, C., FALCKE, M. et al. (2023). SLAM-drop-seq reveals mRNA kinetic rates throughout the cell cycle. *Mol. Syst. Biol*. **19** (10).

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*. **15** 550.

MARCHESE, F. P., RAIMONDI, I. and HUARTE, M. (2017). The multidimensional mechanisms of long noncoding rna function. *Genome Biol*. **18** 206.

MARIN, J.-M., MENGERSEN, K. and ROBERT, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Bayesian Thinking*: *Modeling and Computation*. *Handbook of Statist*. **25** 459–507. Elsevier, Amsterdam. MR2490536 https://doi.org/10.1016/S0169-7161(05)25016-2

MASTRANTONIO, G., BIBBONA, E. and FURLAN, M. (2024). Supplement to "Multiple latent clustering model for the inference of RNA life-cycle kinetic rates from sequencing data." https://doi.org/10.1214/24-AOAS1945SUPP

MICHEL, M., DEMEL, C., ZACHER, B., SCHWALB, B., KREBS, S., BLUM, H., GAGNEUR, J. and CRAMER, P. (2017). Tt-seq captures enhancer landscapes immediately after t-cell stimulation. *Mol. Syst. Biol*. **13**.

MILLER, C., SCHWALB, B., MAIER, K., SCHULZ, D., DÜMCKE, S., ZACHER, B., MAYER, A., SYDOW, J., MARCINOWSKI, L. et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol*. **7** 458.

MULLER, P. and ROSNER, G. L. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Amer. Statist. Assoc*. **92** 1279–1292.

MURTAGH, F. and LEGENDRE, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classification* **31** 274–295. MR3277707 https://doi.org/10.1007/s00357-014-9161-z

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist*. **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169–186. MR2409721 https://doi.org/10.1093/biomet/asm086

PAPASTAMOULIS, P. and RATTRAY, M. (2018). A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **67** 3–23. MR3758753 https://doi.org/10.1111/rssc.12213

RABANI, M., LEVIN, J. Z., FAN, L., ADICONIS, X., RAYCHOWDHURY, R., GARBER, M., GNIRKE, A., NUSBAUM, C., HACOHEN, N. et al. (2011). Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nat. Biotechnol*. **29** 436–442.

RABANI, M., RAYCHOWDHURY, R., JOVANOVIC, M., ROONEY, M., STUMPO, D. J., PAULI, A., HACOHEN, N., SCHIER, A. F., BLACKSHEAR, P. J. et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic rna regulatory strategies. *Cell* **159** 1698–1710.

RUGGERO, D. (2009). The role of myc-induced protein synthesis in cancer. *Cancer Res*. **69** 8839–8843.

RUMMEL, T., SAKELLARIDI, L. and ERHARD, F. (2023). grandr: A comprehensive package for nucleotide conversion rna-seq data analysis. *Nat. Commun*. **14** 3559.

SCHOFIELD, J. A., DUFFY, E. E., KIEFER, L., SULLIVAN, M. C. and SIMON, M. D. (2018). Timelapse-seq: Adding a temporal dimension to rna sequencing through nucleoside recoding. *Nat. Methods* **15** 221–225.

SCHWALB, B., SCHULZ, D., SUN, M., ZACHER, B., DÜMCKE, S., MARTIN, D. E., CRAMER, P. and TRESCH, A. (2012). Measurement of genome-wide rna synthesis and decay rates with dynamic transcriptome analysis (dta). *Bioinformatics* **28** 884–885.

SLACK, F. J. and CHINNAIYAN, A. M. (2019). The role of non-coding rnas in oncology. *Cell* **179** 1033–1055.

STINE, Z. E., WALTON, Z. E., ALTMAN, B. J., HSIEH, A. L. and DANG, C. V. (2015). Myc, metabolism, and cancer. *Cancer Discov.* **5** 1024–1039.

SUN, S., HOOD, M., SCOTT, L., PENG, Q., MUKHERJEE, S., TUNG, J. and ZHOU, X. (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Res*. **45** e106–e106.

TAN, J. Y., BIASINI, A., YOUNG, R. S. and MARQUES, A. C. (2020). Splicing of enhancer-associated lincrnas contributes to enhancer activity. *Life Sci. Alliance* **3** e202000663.

TATARINOVA, T., NEELY, M., BARTROFF, J., VAN GUILDER, M., YAMADA, W., BAYARD, D., JELLIFFE, R., LEARY, R., CHUBATIUK, A. et al. (2013). Two general methods for population pharmacokinetic modeling: Non-parametric adaptive grid and non-parametric Bayesian. *J. Pharmacokinet. Pharmacodyn*. **40** 189–199.

THE GENE ONTOLOGY CONSORTIUM (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. **47** D330–D338. https://doi.org/10.1093/nar/gky1055

TIBERI, S. and ROBINSON, M. D. (2020). Bandits: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty. *Genome Biol*. **21**.

TUERK, A., WIKTORIN, G. and GÜLER, S. (2017). Mixture models reveal multiple positional bias types in rna-seq data and lead to accurate transcript concentration estimates. *PLoS Comput. Biol*. **13** 1–25.

UVAROVSKII, A. and DIETERICH, C. (2017). pulser: Versatile computational analysis of rna turnover from metabolic labeling experiments. *Bioinformatics* **33** 3305–3307.

VANDEVENNE, M., DELMARCELLE, M. and GALLENI, M. (2019). RNA regulatory networks as a control of stochasticity in biological systems. *Front. Genet*. **10** 403. https://doi.org/10.3389/fgene.2019.00403

WACHUTKA, L., CAIZZI, L., GAGNEUR, J. and CRAMER, P. (2019). Global donor and acceptor splicing site kinetics in human cells. *eLife* **8**. https://doi.org/10.7554/eLife.45056

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

WALKER, S. and WAKEFIELD, J. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya*, *Ser. B* **60** 196–214. MR1717082

YU, G., WANG, L.-G., HAN, Y. and HE, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics*. *J. Integr. Biol*. **16** 284–287.

ZEISEL, A., KÖSTLER, W. J., MOLOTSKI, N., TSAI, J. M., KRAUTHGAMER, R., JACOB-HIRSCH, J., RECHAVI, G., SOEN, Y., JUNG, S. et al. (2011). Coupled pre-mrna and mrna dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol*. *Syst*. *Biol*. **7**.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3486–3506
https://doi.org/10.1214/24-AOAS1947

PREDICTING MILK TRAITS FROM SPECTRAL DATA USING BAYESIAN PROBABILISTIC PARTIAL LEAST SQUARES REGRESSION

BY SZYMON URBAS[1,a], PIERRE LOVERA[2,c], ROBERT DALY[2,d], ALAN O'RIORDAN[2,e], DONAGH BERRY[3,f] AND ISOBEL CLAIRE GORMLEY[1,b]

[1]*School of Mathematics & Statistics, University College Dublin,* [a]*szymongurbas@gmail.com,* [b]*claire.gormley@ucd.ie*
[2]*Tyndall National Institute, University College Cork,* [c]*pierre.lovera@tyndall.ie,* [d]*robert.daly@tyndall.ie,* [e]*alan.oriordan@tyndall.ie*
[3]*Animal & Grassland Research and Innovation Centre, Teagasc,* [f]*donagh.berry@teagasc.ie*

High-dimensional spectral data—routinely generated in dairy production—are used to predict a range of traits in milk products. Partial least squares (PLS) regression is ubiquitously used for these prediction tasks. However, PLS regression is not typically viewed as arising from a probabilistic model, and parameter uncertainty is rarely quantified. Additionally, PLS regression does not easily lend itself to model-based modifications, coherent prediction intervals are not readily available, and the process of choosing the latent-space dimension, Q, can be subjective and sensitive to data size.

We introduce a Bayesian latent-variable model, emulating the desirable properties of PLS regression while accounting for parameter uncertainty in prediction. The need to choose Q is eschewed through a nonparametric shrinkage prior. The flexibility of the proposed Bayesian partial least squares (BPLS) regression framework is exemplified by considering sparsity modifications and allowing for multivariate response prediction.

The BPLS regression framework is used in two motivating settings: (1) multivariate trait prediction from mid-infrared spectral analyses of milk samples and (2) milk pH prediction from surface-enhanced Raman spectral data. The prediction performance of BPLS regression at least matches that of PLS regression. Additionally, the provision of correctly calibrated prediction intervals objectively provides richer, more informative inference for stakeholders in dairy production.

REFERENCES

AERNOUTS, B., POLSHIN, E., LAMMERTYN, J. and SAEYS, W. (2011). Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: Reflectance or transmittance? *J. Dairy Sci.* **94** 5315–5329. https://doi.org/10.3168/jds.2011-4354

BARKER, M. and RAYENS, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*: *A Journal of the Chemometrics Society* **17** 166–173.

BEEBE, K. R., PELL, R. J. and SEASHOLTZ, M. B. (1998). *Chemometrics*: *A Practical Guide*. Wiley, New York; Chichester.

BEHKAMI, S., ZAIN, S. M., GHOLAMI, M. and KHIR, M. F. A. (2019). Classification of cow milk using artificial neural network developed from the spectral data of single- and three-detector spectrophotometers. *Food Chem.* **294** 309–315. https://doi.org/10.1016/j.foodchem.2019.05.060

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

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

BONFATTI, V., DI MARTINO, G. and CARNIER, P. (2011). Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. *J. Dairy Sci.* **94** 5776–5785. https://doi.org/10.3168/jds.2011-4401

Key words and phrases. Milk trait prediction, spectral data, partial least squares, Bayesian factor analysis, high-dimensional statistics.

BOULESTEIX, A.-L. and STRIMMER, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* **8** 32–44. https://doi.org/10.1093/bib/bbl016

BRAND, W., WELLS, A. T., SMITH, S. L., DENHOLM, S. J., WALL, E. and COFFEY, M. P. (2021). Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. *J. Dairy Sci.* **104** 4980–4990. https://doi.org/10.3168/jds.2020-18367

CHOI, J., ZOU, H. and OEHLERT, G. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Stat. Interface* **3** 429–436. MR2754740 https://doi.org/10.4310/SII.2010.v3.n4.a1

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241 https://doi.org/10.1111/j.1467-9868.2009.00723.x

CHUNG, D., CHUN, H. and KELES, S. (2019). spls: Sparse Partial Least Squares (SPLS) regression and classification. R package version 2.2-3.

COPPA, M., MARTIN, B., HULIN, S., GUILLEMIN, J., GAUZENTES, J. V., PECOU, A. and ANDUEZA, D. (2021). Prediction of indicators of cow diet composition and authentication of feeding specifications of protected designation of origin cheese using mid-infrared spectroscopy on milk. *J. Dairy Sci.* **104** 112–125. https://doi.org/10.3168/jds.2020-18468

COPPA, M., VANLIERDE, A., BOUCHON, M., JURQUET, J., MUSATI, M., DEHARENG, F. and MARTIN, C. (2022). Methodological guidelines: Cow milk mid-infrared spectra to predict reference enteric methane data collected by an automated head-chamber system. *J. Dairy Sci.* **105** 9271–9285. https://doi.org/10.3168/jds.2022-21890

DE JONG, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **18** 251–263.

DE MARCHI, M., TOFFANIN, V., CASSANDRO, M. and PENASA, M. (2014). Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits1. *J. Dairy Sci.* **97** 1171–1186. https://doi.org/10.3168/jds.2013-6799

DENHOLM, S., BRAND, W., MITCHELL, A., WELLS, A., KRZYZELEWSKI, T., SMITH, S., WALL, E. and COFFEY, M. (2020). Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. *J. Dairy Sci.* **103** 9355–9367.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. MR0829555 https://doi.org/10.1214/aos/1176349830

DUMPLER, J., HUPPERTZ, T. and KULOZIK, U. (2020). Invited review: Heat stability of milk and concentrated milk: Past, present, and future research objectives. *J. Dairy Sci.* **103** 10986–11007. https://doi.org/10.3168/jds.2020-18605

DURANTE, D. (2017). A note on the multiplicative gamma process. *Statist. Probab. Lett.* **122** 198–204. MR3584158 https://doi.org/10.1016/j.spl.2016.11.014

EL BOUHADDANI, S., UH, H.-W., HAYWARD, C., JONGBLOED, G. and HOUWING-DUISTERMAAT, J. (2018). Probabilistic partial least squares model: Identifiability, estimation and application. *J. Multivariate Anal.* **167** 331–346. MR3830650 https://doi.org/10.1016/j.jmva.2018.05.009

EL BOUHADDANI, S., UH, H.-W., JONGBLOED, G. and HOUWING-DUISTERMAAT, J. (2022). Statistical integration of heterogeneous omics data: Probabilistic two-way partial least squares (PO2PLS). *J. R. Stat. Soc. Ser. C. Appl. Stat.* **71** 1451–1470. MR4511118 https://doi.org/10.1111/rssc.12583

FARRELL, H. M. J., JIMENEZ-FLORES, R., BLECK, G. T., BROWN, E. M., BUTLER, J. E., CREAMER, L. K., HICKS, C. L., HOLLAR, C. M., NG-KWAI-HANG, K. F. et al. (2004). Nomenclature of the proteins of cows' milk—sixth revision. *J. Dairy Sci.* **87** 1641–1674.

FILZMOSER, P., GSCHWANDTNER, M. and TODOROV, V. (2012). Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom.* **26** 42–51.

FILZMOSER, P. and VARMUZA, K. (2017). Chemometrics: Multivariate statistical analysis in chemometrics. R package version 1.4.2.

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135. https://doi.org/10.1080/00401706.1993.10485033

FRIZZARIN, M., GORMLEY, I., BERRY, D., MURPHY, T., CASA, A., LYNCH, A. and MCPARLAND, S. (2021). Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods. *J. Dairy Sci.* **104** 7438–7447.

FRIZZARIN, M., GORMLEY, I. C., BERRY, D. P. and MCPARLAND, S. (2023). Estimation of body condition score change in dairy cows in a seasonal calving pasture-based system using routinely available milk mid-infrared spectra and machine learning techniques. *J. Dairy Sci.* **106** 4232–4244. https://doi.org/10.3168/jds.2022-22394

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

FRUHWIRTH-SCHNATTER, S. and LOPES, H. F. (2014). Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical Report No. 345. Insper Working Paper.

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

GWEE, X. Y., GORMLEY, I. C. and FOP, M. (2023). A latent shrinkage position model for binary and count network data. *Bayesian Anal.* 1–29. https://doi.org/10.1214/23-BA1403

HANSEN, B., AVALOS-PACHECO, A., RUSSO, M. and VITO, R. D. (2023). Fast variational inference for Bayesian factor analysis in single and multi-study settings.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12** 69–82.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

HUBERT, M. and BRANDEN, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*: *A Journal of the Chemometrics Society* **17** 537–549.

KANDEEL, S. A., MEGAHED, A. A., EBEID, M. H. and CONSTABLE, P. D. (2019). Ability of milk pH to predict subclinical mastitis and intramammary infection in quarters from lactating dairy cattle. *J. Dairy Sci.* **102** 1417–1427. https://doi.org/10.3168/jds.2018-14993

KOURTI, T. (2002). Process analysis and abnormal situation detection: From theory to practice. *IEEE Control Syst. Mag.* **22** 10–25.

LI, S., GAO, J., NYAGILO, J. O. and DAVE, D. P. (2010). Eigenspectra, a robust regression method for multiplexed Raman spectra analysis. In 2010 *IEEE International Conference on Bioinformatics and Biomedicine* (*BIBM*) 525–530. IEEE.

LIEBMANN, B., FRIEDL, A. and VARMUZA, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Anal. Chim. Acta* **642** 171–178. https://doi.org/10.1016/j.aca.2008.10.069

LILAND, K. H., MEVIK, B.-H. and WEHRENS, R. (2022). pls: Partial least squares and principal component regression. R package version 2.8-1.

LINDGREN, F., GELADI, P. and WOLD, S. (1993). The kernel algorithm for PLS. *J. Chemom.* **7** 45–59.

MCDERMOTT, A., VISENTIN, G., MARCHI, M. D., BERRY, D. P., FENELON, M. A., O'CONNOR, P. M., KENNY, O. A. and MCPARLAND, S. (2016). Prediction of individual milk proteins including free amino acids in bovine milk using mid-infrared spectroscopy and their correlations with milk processing characteristics. *J. Dairy Sci.* **99** 3171–3182. https://doi.org/10.3168/jds.2015-9747

MCPARLAND, D., PHILLIPS, C. M., BRENNAN, L., ROCHE, H. M. and GORMLEY, I. C. (2017). Clustering high-dimensional mixed data to uncover sub-phenotypes: Joint analysis of phenotypic and genotypic data. *Stat. Med.* **36** 4548–4569. MR3731239 https://doi.org/10.1002/sim.7371

MURPHY, K., VIROLI, C. and GORMLEY, I. C. (2020). Infinite mixtures of infinite factor analysers. *Bayesian Anal.* **15** 937–963. MR4132655 https://doi.org/10.1214/19-BA1179

NAIK, P. and TSAI, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 763–771. MR1796290 https://doi.org/10.1111/1467-9868.00262

OVASKAINEN, O., ABREGO, N., HALME, P. and DUNSON, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7** 549–555.

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

PRESS, S. J. and SHIGEMASU, K. (1989). Bayesian inference in factor analysis. In *Contributions to Probability and Statistics* 271–287. Springer, New York. MR1024336

ROCKOVA, V. (2023). Adaptive Bayesian predictive inference. Preprint. Available at arXiv:2309.02369.

ROUSSEAU, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annu. Rev. Stat. Appl.* **3** 211–231.

SCHALM, O. W. and NOORLANDER, D. O. (1957). Experiments and observations leading to development of the California mastitis test. *J. Am. Vet. Med. Assoc.* **130** 199–204.

SONG, X.-Y. and LEE, S.-Y. (2001). Bayesian estimation and test for factor analysis model with continuous and polytomous data in several populations. *Br. J. Math. Stat. Psychol.* **54** 237–263.

R CORE TEAM (2023). *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

TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 611–622. MR1707864 https://doi.org/10.1111/1467-9868.00196

TOLEDO-ALVARADO, H., PÉREZ-CABAL, M. A., TEMPELMAN, R. J., CECCHINATO, A., BITTANTE, G., DE LOS CAMPOS, G. and VAZQUEZ, A. I. (2021). Association between days open and milk spectral data in dairy cows. *J. Dairy Sci.* **104** 3665–3675. https://doi.org/10.3168/jds.2020-19031

TRYGG, J. and WOLD, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*: *A Journal of the Chemometrics Society* **16** 119–128.

TRYGG, J. and WOLD, S. (2003). O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* **17** 53–64.

URBAS, S., LOVERA, P., DALY, R., O'RIORDAN, A., BERRY, D. and GORMLEY, I. C (2024). Supplement to "Predicting milk traits from spectral data using Bayesian probabilistic partial least squares regression." https://doi.org/10.1214/24-AOAS1947SUPPA, https://doi.org/10.1214/24-AOAS1947SUPPB

VAN DER VOET, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.* **25** 313–323.

VATS, D. and KNUDSON, C. (2021). Revisiting the Gelman-Rubin diagnostic. *Statist. Sci.* **36** 518–529. MR4323050 https://doi.org/10.1214/20-sts812

VIDAURRE, D., VAN GERVEN, M. A. J., BIELZA, C., LARRAÑAGA, P. and HESKES, T. (2013). Bayesian sparse partial least squares. *Neural Comput.* **25** 3318–3339. MR3154315 https://doi.org/10.1162/NECO_a_00524

VISENTIN, G., MCDERMOTT, A., MCPARLAND, S., BERRY, D. P., KENNY, O. A., BRODKORB, A., FENELON, M. A. and MARCHI, M. D. (2015). Prediction of bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows. *J. Dairy Sci.* **98** 6620–6629. https://doi.org/10.3168/jds.2015-9323

WILLIAMS, A., FLYNN, K. J., XIA, Z. and DUNSTAN, P. R. (2016). Multivariate spectral analysis of pH SERS probes for improved sensing capabilities. *J. Raman Spectrosc.* **47** 819–827.

WOLD, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: Some current developments. In *Multivariate Analysis*, *III* (*Proc. Third Internat. Sympos.*, *Wright State Univ.*, *Dayton*, *Ohio*, 1972) 383–407. Academic Press, New York. MR0343487

ZHENG, J., SONG, Z. and GE, Z. (2016). Probabilistic learning of partial least squares regression model: Theory and industrial applications. *Chemom. Intell. Lab. Syst.* **158** 80–90.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3507–3527
https://doi.org/10.1214/24-AOAS1949

A NEW DESIGN FOR OBSERVATIONAL STUDIES APPLIED TO THE STUDY OF THE EFFECTS OF HIGH SCHOOL FOOTBALL ON COGNITION LATE IN LIFE

BY KATHERINE BRUMBERG[1,a], DYLAN S. SMALL[2,b] AND PAUL R. ROSENBAUM[2,c]

[1]*Department of Statistics, University of Michigan,* [a]*kbrum@umich.edu*
[2]*Department of Statistics and Data Science, Wharton School, University of Pennsylvania,* [b]*dsmall@wharton.upenn.edu,* [c]*rosenbaum@wharton.upenn.edu*

Do the impacts that occur when playing high school football have concussive effects that accelerate cognitive decline late in life? We examine this possibility using newly available cognitive data describing people in 2020 who graduated high school in 1957. Someone who was 18 in 1957 would be 81 in 2020. For this comparison we develop a new design for an observational study, called a triples design, and discuss its advantages and construction. A triples design consists of M blocks of size 3, where a block contains either one treated individual and two controls or two treated individuals and one control. A triples design is the simplest design that uses weights, with just two weights. The "entire number" is $\{1 - e(\mathbf{x})\}/e(\mathbf{x})$, where $e(\mathbf{x})$ is the propensity score at covariate $\mathbf{x}$, so it is the ratio of controls-to-treated expected at $\mathbf{x}$. Unlike a matched pairs design, which can remove the bias from observed covariates when the "entire number" exceeds 1, the triples design can succeed when the entire number exceeds 1/2, reflecting the possibility of matching two treated individuals to the same control. Like full matching, a triples design can match more people than can matched pairs, yet have smaller within-block covariate distances. Unlike full matching, there are no matched pairs. Like matching with multiple controls, a triples design will have a larger design sensitivity than a design which includes matched pairs, under simple models for continuous outcomes; that is, in favorable situations the design is expected to report greater insensitivity to unmeasured biases. Because there are just two weights, it is easy to construct weighted graphics for exploratory displays from triples designs. A heuristic algorithm containing network optimization constructs the design.

REFERENCES

ALOSCO, M. L., MEZ, J., TRIPODIS, Y. et al. (2018). Age of first exposure to tackle football and chronic traumatic encephalopathy. *Ann. Neurol.* **83** 886–901.

BAUER, P. and KIESER, M. (1996). A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika* **83** 934–937. MR1440056 https://doi.org/10.1093/biomet/83.4.934

BERTSEKAS, D. P. and TSENG, P. (1988). The RELAX codes for linear minimum cost network flow problems. *Ann. Oper. Res.* **13** 125–190. MR0950991 https://doi.org/10.1007/BF02288322

BRUMBERG, K., ELLIS, D. E., SMALL, D. S., HENNESSY, S. and ROSENBAUM, P. R. (2023). Using natural strata when examining unmeasured biases in an observational study of neurological side effects of antibiotics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **72** 314–329. MR4719278 https://doi.org/10.1093/jrsssc/qlad010

BRUMBERG, K., SMALL, D. S. and ROSENBAUM, P. R. (2024). Supplement to "A new design for observational studies applied to the study of the effects of high school football on cognition late in life." https://doi.org/10.1214/24-AOAS1949SUPPA, https://doi.org/10.1214/24-AOAS1949SUPPB, https://doi.org/10.1214/24-AOAS1949SUPPC

CRAMA, Y. and SPIEKSMA, F. C. (1992). Approximation algorithms for three-dimensional assignment problems with triangle inequalities. *European J. Oper. Res.* **60** 273–279.

Key words and phrases. Aberrant effects, causal inference, design sensitivity, full matching, propensity score, subset matching, triples design.

DESHPANDE, S. K., HASEGAWA, R. B., RABINOWITZ, A. R., WHYTE, J., ROAN, C. L., TABATABAEI, A., BAIOCCHI, M., KARLAWISH, J. H., MASTER, C. L. et al. (2017). Association of playing high school football with cognition and mental health later in life. *JAMA Neurol*. **74** 909–918. https://doi.org/10.1001/jamaneurol.2017.1317

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc*. **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist*. **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

HOWARD, S. R. and PIMENTEL, S. D. (2021). The uniform general signed rank test and its design sensitivity. *Biometrika* **108** 381–396. MR4259138 https://doi.org/10.1093/biomet/asaa072

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

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2019). Using approximation algorithms to build evidence factors and related designs for observational studies. *J. Comput. Graph. Statist*. **28** 698–709. MR4007751 https://doi.org/10.1080/10618600.2019.1584900

KNOPMAN, D. S., ROBERTS, R. O., GEDA, Y. E., PANKRATZ, V. S., CHRISTIANSON, T. J. H., PETERSEN, R. C. and ROCCA, W. A. (2010). Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology* **34** 34–42. https://doi.org/10.1159/000255464

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

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161 https://doi.org/10.1093/biomet/66.1.163

MEZ, J., DANESHVAR, D. H., KIERNAN, P. T. et al. (2017). Clinicopathological evaluation of chronic traumatic encephalopathy in players of American football. *J. Amer. Med. Assoc*. **318** 360–370.

NATIONAL FEDERATION OF STATE HIGH SCHOOL ASSOCIATIONS (2022). NFHS handbook 2022-3.

NATTINO, G., LU, B., SHI, J., LEMESHOW, S. and XIANG, H. (2021). Triplet matching for estimating causal effects with three treatment arms: A comparative study of mortality by trauma center level. *J. Amer. Statist. Assoc*. **116** 44–53. MR4227673 https://doi.org/10.1080/01621459.2020.1737078

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. (Reprint in English of Neyman (1923). *Statist. Sci*. 465–472.

PIMENTEL, S. D. (2016). Large, sparse optimal matching with R package rebalance. *Obs. Stud*. **2** 4–23.

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med*. **34** 4070–4082. MR3431322 https://doi.org/10.1002/sim.6593

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915 https://doi.org/10.1093/biomet/74.1.13

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

ROSENBAUM, P. R. (1994). Coherence in observational studies. *Biometrics* **50** 368–374.

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

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist*. **59** 147–152. MR2133562 https://doi.org/10.1198/000313005X42831

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

ROSENBAUM, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist*. **21** 57–71. MR2913356 https://doi.org/10.1198/jcgs.2011.09219

ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118–127. MR3058058 https://doi.org/10.1111/j.1541-0420.2012.01821.x

ROSENBAUM, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc*. **109** 1145–1158. MR3265687 https://doi.org/10.1080/01621459.2013.879261

ROSENBAUM, P. R. (2017). *Observation and Experiment*: *An Introduction to Causal Inference*. Harvard Univ. Press, Cambridge, MA. MR3702029

ROSENBAUM, P. R. (2020). Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl*. **7** 143–176. MR4104189 https://doi.org/10.1146/annurev-statistics-031219-041058

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

ROSENBAUM, P. R. (2024). Bahadur efficiency of observational block designs. *J. Amer. Statist. Assoc*. https://doi.org/10.1080/01621459.2023.2221402.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc*. **79** 516–524.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41** 103–116. MR0793436 https://doi.org/10.2307/2530647

ROSENBAUM, P. R. and RUBIN, D. B. (2023). Propensity scores in the design of observational studies for causal effects. *Biometrika* **110** 1–13. MR4565440 https://doi.org/10.1093/biomet/asac054

ROSENBAUM, P. R. and SILBER, J. H. (2008). Aberrant effects of treatment. *J. Amer. Statist. Assoc*. **103** 240–247. MR2420230 https://doi.org/10.1198/016214507000001274

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.

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

WLS (2022). Wisconsin Longitudinal Study Codebooks: 2020 ILIAD Grad. Available at www.ssc.wisc.edu. Accessed 8/23/23.

ZHAO, Q. (2019). On sensitivity value of pair-matched observational studies. *J. Amer. Statist. Assoc*. **114** 713–722. MR3963174 https://doi.org/10.1080/01621459.2018.1429277

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3528–3548
https://doi.org/10.1214/24-AOAS1950

SCALABLE TEST OF STATISTICAL SIGNIFICANCE FOR PROTEIN-DNA BINDING CHANGES WITH INSERTION AND DELETION OF BASES IN THE GENOME

BY QINYI ZHOU[1,a], CHANDLER ZUO[1,b], YUANNYU ZHANG[2,d], MIN CHEN[1,c], JIAN XU[2,e] AND SUNYOUNG SHIN[3,f]

[1]*Department of Mathematical Sciences, University of Texas at Dallas,* [a]*zhouqy0531@aliyun.com,* [b]*chandler.c.zuo@gmail.com,* [c]*yuannyu.zhang@stjude.org*

[2]*Center of Excellence for Leukemia Studies, Department of Pathology, St. Jude Children's Research Hospital,* [d]*mchen@utdallas.edu,* [e]*jian.xu@stjude.org*

[3]*Department of Mathematics, Pohang University of Science and Technology,* [f]*sunyoungshin@postech.ac.kr*

Mutations in the noncoding DNA, which represents approximately 99% of the human genome, have been crucial to understanding disease mechanisms through dysregulation of disease-associated genes. One key element in gene regulation that noncoding mutations mediate is the binding of proteins to DNA sequences. Insertion and deletion of bases (InDels) are the second most common type of mutations, following single nucleotide polymorphisms, that may impact protein-DNA binding. However, no existing methods can estimate and test the effects of InDels on the process of protein-DNA binding. We develop a novel test of statistical significance, namely, the binding change test (BC test), using a Markov model to evaluate the impact and identify InDels altering protein-DNA binding. The test predicts binding changer InDels of regulatory significance with an efficient importance sampling algorithm generating background sequences in favor of large binding affinity changes. Simulation studies demonstrate its excellent performance. The application to human leukemia data uncovers, in critical cis-regulatory elements, candidate pathological InDels on modulating TF binding in leukemic patients. We develop an R package atIndel, which is available on GitHub.

REFERENCES

AVERY, P. J. (1987). The analysis of intron data and their use in the detection of short signals. *J. Mol. Evol.* **26** 335–340. https://doi.org/10.1007/BF02101152

AVERY, P. J. and HENDERSON, D. A. (1999). Detecting a changed segment in DNA sequences. *J. R. Stat. Soc., Ser. C* **48** 489–503. MR1721441 https://doi.org/10.1111/1467-9876.00167

BAILEY, T. L., WILLIAMS, N., MISLEH, C. and LI, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34** W369–W373.

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

BORODOVSKY, M. and MCININCH, J. (1993). GENMARK: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17** 123–133.

BORODOVSKY, M., MCLNINCH, J. D., KOONIN, E. V., RUDD, K. E., MÉDIGUE, C. and DANCHIN, A. (1995). Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* **23** 3554–3562.

BRAUN, B. S., TUVESON, D. A., KONG, N., LE, D. T., KOGAN, S. C., ROZMUS, J., LE BEAU, M. M., JACKS, T. E. and SHANNON, K. M. (2004). Somatic activation of oncogenic Kras in hematopoietic cells initiates a rapidly fatal myeloproliferative disorder. *Proc. Natl. Acad. Sci. USA* **101** 597–602.

CAPPELLO, L. and PALACIOS, J. A. (2020). Sequential importance sampling for multiresolution Kingman-Tajima coalescent counting. *Ann. Appl. Stat.* **14** 727–751. MR4117827 https://doi.org/10.1214/19-AOAS1313

CHAN, H. P. and ZHANG, N. R. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc.* **102** 595–602. MR2370856 https://doi.org/10.1198/016214506000001392

Key words and phrases. Noncoding mutations, p-value based test statistic, importance sampling, sequence-based models, transcription factor binding, test of significance.

CHAN, H. P., ZHANG, N. R. and CHEN, L. H. Y. (2010). Importance sampling of word patterns in DNA and protein sequences. *J. Comput. Biol.* **17** 1697–1709. MR2749757 https://doi.org/10.1089/cmb.2008.0233

CINGOLANI, P., PLATTS, A., WANG, L. L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. and RUDEN, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* (*Austin*) **6** 80–92. https://doi.org/10.4161/fly.19695

COETZEE, S. G., COETZEE, G. A. and HAZELETT, D. J. (2015). motifbreakR: An R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31** 3847–3849.

THE ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74. https://doi.org/10.1038/nature11247

THE ENCODE PROJECT CONSORTIUM, MOORE, J. E., PURCARO, M. J., PRATT, H. E., EPSTEIN, C. B., SHORESH, N., ADRIAN, J., KAWLI, T., DAVIS, C. A. et al. (2020a). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583** 699–710.

THE ENCODE PROJECT CONSORTIUM, SNYDER, M. P., GINGERAS, T. R., MOORE, J. E., WENG, Z., GERSTEIN, M. B., REN, B., HARDISON, R. C., STAMATOYANNOPOULOS, J. A. et al. (2020b). Perspectives on ENCODE. *Nature* **583** 693–698.

COWAN, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Probab.* **28** 886–892. MR1133796 https://doi.org/10.2307/3214691

HUANG, H., KAO, M.-C. J., ZHOU, X., LIU, J. S. and WONG, W. H. (2004). Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. *J. Comput. Biol.* **11** 1–14.

DELCHER, A. L., HARMON, D., KASIF, S., WHITE, O. and SALZBERG, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27** 4636–4641. https://doi.org/10.1093/nar/27.23.4636

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

FANG, H., BERGMANN, E. A., ARORA, K., VACIC, V., ZODY, M. C., IOSSIFOV, I., O'RAWE, J. A., WU, Y., BARRON, L. T. J. et al. (2016). Indel variant analysis of short-read sequencing data with Scalpel. *Nat. Protoc.* **11** 2529–2548.

FREDRIKSSON, N. J., NY, L., NILSSON, J. A. and LARSSON, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46** 1258–1263.

GRANT, C. E., BAILEY, T. L. and NOBEL, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **7** 1017. https://doi.org/10.1093/bioinformatics/btr064

GUPTA, M. and IBRAHIM, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Amer. Statist. Assoc.* **102** 867–880. MR2411650 https://doi.org/10.1198/016214507000000068

HERTZ, G. Z. and STORMO, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15** 563–577. https://doi.org/10.1093/bioinformatics/15.7.563

HUANG, J., LIU, X., LI, D., SHAO, Z., CAO, H., ZHANG, Y., TROMPOUKI, E., BOWMAN, T. V., ZON, L. I. et al. (2016). Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Dev. Cell* **36** 9–23. https://doi.org/10.1016/j.devcel.2015.12.014

JENSEN, S. T. and LIU, J. S. (2008). Bayesian clustering of transcription factor binding motifs. *J. Amer. Statist. Assoc.* **103** 188–200. MR2420226 https://doi.org/10.1198/016214507000000365

KHAN, A., FORNES, O., STIGLIANI, A., GHEORGHE, M., CASTRO-MONDRAGON, J. A., VAN DER LEE, R., BESSY, A., CHÈNEBY, J., KULKARNI, S. R. et al. (2017). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46** D260–D266. https://doi.org/10.1093/nar/gkx1126

KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J. et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330.

LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615 https://doi.org/10.1198/jasa.2010.tm08177

LI, K., ZHANG, Y., LIU, X., LIU, Y., GU, Z., CAO, H., DICKERSON, K. E., CHEN, M., CHEN, W. et al. (2020). Noncoding variants connect enhancer dysregulation with nuclear receptor signaling in hematopoietic malignancies. *Cancer Discov.* **10** 724–745. https://doi.org/10.1158/2159-8290.CD-19-1128

LI, S., KO, Y. M. and BYON, E. (2021). Nonparametric importance sampling for wind turbine reliability analysis with stochastic computer models. *Ann. Appl. Stat.* **15** 1850–1871. MR4355079 https://doi.org/10.1214/21-aoas1490

LIANG, F. (2002). Dynamically weighted importance sampling in Monte Carlo computation. *J. Amer. Statist. Assoc.* **97** 807–821. MR1941411 https://doi.org/10.1198/016214502388618618

LIN, M., WHITMIRE, S., CHEN, J., FARREL, A., SHI, X. and GUO, J.-T. (2017). Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **7** 1–9.

LIU, X., CHEN, Y., ZHANG, Y., LIU, Y., LIU, N., BOTTEN, G. A., CAO, H., ORKIN, S. H., ZHANG, M. Q. et al. (2020). Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. *Genome Biol.* **21** 1–20.

LIU, X., ZHANG, Y., CHEN, Y., LI, M., ZHOU, F., LI, K., CAO, H., NI, M., LIU, Y. et al. (2017). In situ capture of chromatin interactions by biotinylated DCas9. *Cell* **170** 1028–1043.e19. https://doi.org/10.1016/j.cell.2017.08.003

MACINTYRE, G., BAILEY, J., HAVIV, I. and KOWALCZYK, A. (2010). is-rSNP: A novel technique for in silico regulatory SNP detection. *Bioinformatics* **26** 524–530.

KAHN, H. MARSHALL, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *J. Oper. Res. Soc. Am.* **1** 263–278.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. and CUNNINGHAM, F. (2016). The ensembl variant effect predictor. *Genome Biol.* **17** 1–14.

MCLEAN, C. Y., BRISTOR, D., HILLER, M., CLARKE, S. L., SCHAAR, B. T., LOWE, C. B., WENGER, A. M. and BEJERANO, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28** 495–501. https://doi.org/10.1038/nbt.1630

MENÉNDEZ, M. L., PARDO, L., PARDO, M. C. and ZOGRAFOS, K. (2011). Testing the order of Markov dependence in DNA sequences. *Methodol. Comput. Appl. Probab.* **13** 59–74. MR2755132 https://doi.org/10.1007/s11009-008-9107-1

MILLS, R. E., LUTTIG, C. T., LARKINS, C. E., BEAUCHAMP, A., TSUI, C., PITTARD, W. S. and DEVINE, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16** 1182–1190. https://doi.org/10.1101/gr.4565806

MUKHERJEE, R., PILLAI, N. S. and LIN, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann. Statist.* **43** 352–381. MR3311863 https://doi.org/10.1214/14-AOS1279

MULLANEY, J. M., MILLS, R. E., PITTARD, W. S. and DEVINE, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19** R131–R136.

PORTALES-CASAMAR, E., THONGJUEA, S., KWON, A. T., ARENILLAS, D., ZHAO, X., VALEN, E., YUSUF, D., LENHARD, B., WASSERMAN, W. W. et al. (2010). JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38** D105–D110. https://doi.org/10.1093/nar/gkp950

REINERT, G., SCHBATH, S. and WATERMAN, M. S. (2000). Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.* **7** 1–46. https://doi.org/10.1089/10665270050081360

RETKUTE, R., TOULOUPOU, P., BASÁÑEZ, M.-G., HOLLINGSWORTH, T. D. and SPENCER, S. E. F. (2021). Integrating geostatistical maps and infectious disease transmission models using adaptive multiple importance sampling. *Ann. Appl. Stat.* **15** 1980–1998. MR4355085 https://doi.org/10.1214/21-aoas1486

RIVA, A. (2012). Large-scale computational identification of regulatory SNPs with rSNP-MAPPER. *BMC Genomics* **13** S7. https://doi.org/10.1186/1471-2164-13-S4-S7

SAUNDERS, C. T., WONG, W. S. W., SWAMY, S., BECQ, J., MURRAY, L. J. and CHEETHAM, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28** 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

SEHN, J. K. (2015). Chapter 9—insertions and deletions (indels). In *Clinical Genomics* (S. Kulkarni and J. Pfeifer, eds.) 129–150. Academic Press, Boston, MA. https://doi.org/10.1016/B978-0-12-404748-8.00009-5

STORMO, G. D., SHNEIDER, T. D., GOLD, L. and EHRENFEUCHT, A. (1982). Use of 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.* **10** 2997–3010.

TANG, Z., KANG, B., LI, C., CHEN, T. and ZHANG, Z. (2019). GEPIA2: An enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* **47** W556–W560.

TOMCZAK, K., CZERWIŃSKA, P. and WIZNEROWICZ, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19** A68–A77.

VAN METER, M. E., DÍAZ-FLORES, E., ARCHARD, J. A., PASSEGUÉ, E., IRISH, J. M., KOTECHA, N., NOLAN, G. P., SHANNON, K. and BRAUN, B. S. (2007). K-RasG12D expression induces hyperproliferation and aberrant signaling in primary hematopoietic stem/progenitor cells. *Blood* **109** 3945–3952.

WANG, K., LI, M. and HAKONARSON, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38** e164–e164.

ZHOU, Q. and LIU, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20** 909–916.

ZHOU, Q., ZUO, C., ZHANG, Y., CHEN, M., XU, J. and SHIN, S. (2024). Supplement to "Scalable test of statistical significance for protein-DNA binding changes with insertion and deletion of bases in the genome." https://doi.org/10.1214/24-AOAS1950SUPPA, https://doi.org/10.1214/24-AOAS1950SUPPB

ZUO, C., SHIN, S. and KELEŞ, S. (2015). atSNP: Transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31** 3353–3355. https://doi.org/10.1093/bioinformatics/btv328

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3549–3569
https://doi.org/10.1214/24-AOAS1952

SPATIO-TEMPORAL ANALYSIS OF DEPENDENT RISK WITH AN APPLICATION TO CYBERATTACKS DATA

BY SONGHYUN KIM[1,a], CHAE YOUNG LIM[1,b] AND YEONWOO RHO[2,c]

[1]*Department of Statistics, Seoul National University,* [a]*thanyou833@gmail.com,* [b]*twinwood@snu.ac.kr*
[2]*Department of Mathematical Sciences, Michigan Technological University,* [c]*yrho@mtu.edu*

Cybersecurity is an important issue given the increasing risks due to cyberattacks in many areas. Cyberattacks could result in huge losses such as data breaches, failures in the control systems of infrastructures, physical damages in manufacturing industries, etc. As a result, cybersecurity-related research has grown rapidly for in-depth analysis. One main interest is to understand the correlated nature of cyberattack data. To understand such characteristics, we propose a spatio-temporal model for the hostwisely aggregated cyberattack data by incorporating the characteristics of the attackers. We develop a new dissimilarity measure as a proxy of spatial distance to be integrated into the model. The proposed model can be considered as a spatial extension of the GARCH model. The estimation is carried out using a Bayesian approach, which is demonstrated to work well in simulations. The proposed model is applied to publicly available honeypot data after the data are divided by selected features of the attackers via clustering. The estimated model parameters vary by groups of attackers, which was not revealed by modeling the entire dataset.

REFERENCES

AGARAP, A. F. M. (2018). A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data. In *Proceedings of the* 2018 10*th International Conference on Machine Learning and Computing* 26–30.

ANIRUDH, M., THILEEBAN, S. A. and NALLATHAMBI, D. J. (2017). Use of honeypots for mitigating DoS attacks targeted on IoT networks. In 2017 *International Conference on Computer, Communication and Signal Processing* (*ICCCSP*) 1–4.

BENEDICT, S. (2023). EA-POT: An explainable AI assisted blockchain framework for honeypot IP predictions. *Acta Cybernet*. **26** 149–173.

BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051 https://doi.org/10.1016/0304-4076(86)90063-1

BOLLERSLEV, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Rev. Econ. Stat*. **72** 498–505.

BOLLERSLEV, T., ENGLE, R. F. and NELSON, D. B. (1994). Chapter 49 arch models. In *Handbook of Econometrics*, *Vol. IV*. *Handbooks in Econom*. **2** 2959–3038. North-Holland, Amsterdam. MR1315984

BOLLERSLEV, T., ENGLE, R. F. and WOOLDRIDGE, J. M. (1988). A capital asset pricing model with time-varying covariances. *J. Polit. Econ*. **96** 116–131.

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

CHEN, C. W. S., WATANABE, T. and LIN, E. M. H. (2023). Bayesian estimation of realized GARCH-type models with application to financial tail risk management. *Econom. Stat*. **28** 30–46. MR4644290 https://doi.org/10.1016/j.ecosta.2021.03.006

CHEN, Y.-Z., HUANG, Z.-G., XU, S. and LAI, Y.-C. (2015). Spatiotemporal patterns and predictability of cyberattacks. *PLoS ONE* **10** e0124472.

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

CHRIST, M., KEMPA-LIEHR, A. W. and FEINDT, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. arXiv preprint. Available at arXiv:1610.07717.

Key words and phrases. Clustering, cyberattack, honeypot data, MCMC, spatial-GARCH model.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. Revised reprint of the 1991 edition, a Wiley-Interscience Publication. MR1239641 https://doi.org/10.1002/9781119115151

DAGON, D., QIN, X., GU, G., LEE, W., GRIZZARD, J., LEVINE, J. and OWEN, H. (2004). Honeystat: Local worm detection using honeypots. In *Recent Advances in Intrusion Detection* (E. Jonsson, A. Valdes and M. Almgren, eds.) 39–58. Springer, Berlin.

DAVID, A. O. and OLUWASOLA, O. O. (2020). Zero day attack prediction with parameter setting using bi direction recurrent neural network in cyber security. *Int. J. Comput. Sci. Inf. Secur.* **18** 111–118.

DICKEY, D. A. and FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74** 427–431. MR0548036

ELING, M. and JUNG, K. (2018). Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance Math. Econom.* **82** 167–180. MR3850616 https://doi.org/10.1016/j.insmatheco.2018.07.003

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

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

FANG, X., XU, M., XU, S. and ZHAO, P. (2019). A deep learning framework for predicting cyber attacks rates. *EURASIP J. Inf. Secur.* **2019** 5.

FANG, Z., XU, M., XU, S. and HU, T. (2021). A framework for predicting data breach risk: Leveraging dependence to cope with sparsity. *IEEE Trans. Inform. Forensics Secur.* **16** 2186–2201.

FENG, C., WANG, H., LU, N., CHEN, T., HE, H., LU, Y. et al. (2014). Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **26** 105–109.

FENG, L. and SHI, Y. (2017). Fractionally integrated GARCH model with tempered stable distribution: A simulation study. *J. Appl. Stat.* **44** 2837–2857. MR3721076 https://doi.org/10.1080/02664763.2016.1266310

FRANCQ, C. and ZAKOÏAN, J.-M. (2010). *GARCH Models*: *Structure*, *Statistical Inference and Financial Applications*. Wiley, Chichester. MR3186556 https://doi.org/10.1002/9780470670057

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

HAFNER, C. M. (2009). Garch modeling. In *Complex Systems in Finance and Econometrics* (R. A. Meyers, ed.) 464–483. Springer, Berlin.

HANSEN, P. R. and LUNDE, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)? *J. Appl. Econometrics* **20** 873–889. MR2223415 https://doi.org/10.1002/jae.800

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

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

HOBBS, A. (2021). The colonial pipeline hack: Exposing vulnerabilities in U.S. cybersecurity. In *SAGE Business Cases*. SAGE Publications.

HØLLELAND, S. and KARLSEN, H. A. (2020). A stationary spatio-temporal Garch model. *J. Time Series Anal.* **41** 177–209. MR4086183 https://doi.org/10.1111/jtsa.12498

HYNDMAN, R. J. and ATHANASOPOULOS, G. (2018). Forecasting: Principles and Practice. OTexts.

JACOBS, J. (2014). DDS dataset collection. DDS Dataset Collection. Available at http://web.archive.org/web/20150108174904/http://datadrivensecurity.info/blog/pages/dds-dataset-collection.html. Archived via Wayback Machine. Available at https://datadrivensecurity.info/blog/pages/dds-dataset-collection.html (Accessed on 01 Jan 2023).

JAVED, F. and MANTALOS, P. (2013). GARCH-type models and performance of information criteria. *Comm. Statist. Simulation Comput.* **42** 1917–1933. MR3042808

KIM, S., LIM, C. Y. and RHO, Y. (2024). Supplement to "Spatio-Temporal Analysis of Dependent Risk with an Application to Cyberattacks Data." https://doi.org/10.1214/24-AOAS1952SUPPA, https://doi.org/10.1214/24-AOAS1952SUPPB

KODINARIYA, T. M. and MAKWANA, P. R. (2013). Review on determining number of cluster in k-means clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **1** 90–95.

KREIBICH, C. and CROWCROFT, J. (2004). Honeycomb: Creating intrusion detection signatures using honeypots. *Comput. Commun. Rev.* **34** 51–56.

KWIATKOWSKI, D., PHILLIPS, P. C., SCHMIDT, P. and SHIN, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econometrics* **54** 159–178.

KWON, D., NATARAJAN, K., SUH, S. C., KIM, H. and KIM, J. (2018). An empirical study on network anomaly detection using convolutional neural networks. In 2018 *IEEE* 38*th International Conference on Distributed Computing Systems* (*ICDCS*) 1595–1598.

LELAND, W. E., TAQQU, M. S., WILLINGER, W. and WILSON, D. V. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw.* **2** 1–15.

LESLIE, N. O., HARANG, R. E., KNACHEL, L. P. and KOTT, A. (2018). Statistical models for the number of successful cyber intrusions. *J. Defense Model. Simul.* **15** 49–63.

LI, Q., TIAN, Y., WU, Q., CAO, Q., SHEN, H. and LONG, H. (2020). A cloud-fog-edge closed-loop feedback security risk prediction method. *IEEE Access* **8** 29004–29020.

LIANG, G., WELLER, S. R., ZHAO, J., LUO, F. and DONG, Z. Y. (2017). The 2015 Ukraine blackout: Implications for false data injection attacks. *IEEE Trans. Power Syst.* **32** 3317–3318.

LING, S. and LI, W. K. (1997). On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *J. Amer. Statist. Assoc.* **92** 1184–1194. MR1482150 https://doi.org/10.2307/2965585

LING, X., RHO, Y. and TEN, C.-W. (2019). Predicting global trend of cybersecurity on continental honeynets using vector autoregression. In 2019 *IEEE PES Innovative Smart Grid Technologies Europe* (*ISGT-Europe*) 1–5.

MARCHESE, M., KYRIAKOU, I., TAMVAKIS, M. and DI IORIO, F. (2020). Forecasting crude oil and refined products volatilities and correlations: New evidence from fractionally integrated multivariate GARCH models. *Energy Econ.* **88** 104757.

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

NELSON, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory* **6** 318–334. MR1085577 https://doi.org/10.1017/S0266466600005296

NURSETYO, A., IGNATIUS MOSES SETIADI, D. R., RACHMAWANTO, E. H. and SARI, C. A. (2019). Website and network security techniques against brute force attacks using honeypot. In 2019 4*th International Conference on Informatics and Computing* (*ICIC*) 1–6.

OTTO, P., SCHMID, W. and GARTHOFF, R. (2018). Generalised spatial and spatiotemporal autoregressive conditional heteroscedasticity. *Spat. Stat.* **26** 125–145. MR3846283 https://doi.org/10.1016/j.spasta.2018.07.005

OTTO, P., SCHMID, W. and GARTHOFF, R. (2021). Stochastic properties of spatial and spatiotemporal ARCH models. *Statist. Papers* **62** 623–638. MR4232910 https://doi.org/10.1007/s00362-019-01106-x

PENG, C., XU, M., XU, S. and HU, T. (2018). Modeling multivariate cybersecurity risks. *J. Appl. Stat.* **45** 2718–2740. MR3861481 https://doi.org/10.1080/02664763.2018.1436701

RANGANATH, M. and KEATING, M. (2022). How to detect suspicious activity in your AWS account by using private decoy resources. AWS Security Blog. Available at https://aws.amazon.com/blogs/security/how-to-detect-suspicious-activity-in-your-aws-account-by-using-private-decoy-resources/ (Accessed on 01 Jan 2023).

SAÏD, S. E. and DICKEY, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **71** 599–607. MR0775407 https://doi.org/10.1093/biomet/71.3.599

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

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423, 623–656. MR0026286 https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SHYLA, S. and BHATNAGAR, V. (2021). The geo-spatial distribution of targeted attacks sources using honeypot networks. In 2021 11*th International Conference on Cloud Computing, Data Science and Engineering* (*Confluence*) 600–604.

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

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

SULLIVAN, J. E. and KAMENSKY, D. (2017). How cyber-attacks in Ukraine show the vulnerability of the U.S. power grid. *Electr. J.* **30** 30–35.

TANG, M., ALAZAB, M. and LUO, Y. (2017). Big data for cybersecurity: Vulnerability disclosure trends and dependencies. *IEEE Trans. Big Data* **5** 317–329.

THONNARD, O. and DACIER, M. (2008). A framework for attack patterns' discovery in honeynet data. *Digit. Investig.* **5** S128–S139.

TSAY, R. S. (2014). *Multivariate Time Series Analysis*: *With R and Financial Applications*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3236787

TSE, Y. K. and TSUI, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *J. Bus. Econom. Statist.* **20** 351–362. MR1939906 https://doi.org/10.1198/073500102288618496

VISHWAKARMA, R. and JAIN, A. K. (2019). A honeypot with machine learning based detection framework for defending IoT based botnet DDoS attacks. In 2019 3*rd International Conference on Trends in Electronics and Informatics* (*ICOEI*) 1019–1024.

WAHAB, O. A., BENTAHAR, J., OTROK, H. and MOURAD, A. (2019). Resource-aware detection and defense system against multi-type attacks in the cloud: Repeated Bayesian Stackelberg game. *IEEE Trans*. *Dependable Secure Comput*. **18** 605–622.

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

XU, M., HUA, L. and XU, S. (2017). A vine copula model for predicting the effectiveness of cyber defense early-warning. *Technometrics* **59** 508–520. MR3740967 https://doi.org/10.1080/00401706.2016.1256841

XU, M., SCHWEITZER, K. M., BATEMAN, R. M. and XU, S. (2018). Modeling and predicting cyber hacking breaches. *IEEE Trans*. *Inform*. *Forensics Secur.* **13** 2856–2871.

ZHAN, Z., XU, M. and XU, S. (2013). Characterizing honeypot-captured cyber attacks: Statistical framework and case study. *IEEE Trans*. *Inform*. *Forensics Secur.* **8** 1775–1789.

ZHAN, Z., XU, M. and XU, S. (2015). Predicting cyber attack rates with extreme values. *IEEE Trans*. *Inform*. *Forensics Secur.* **10** 1666–1677. MR3389187

ZHANG WU, M., LUO, J., FANG, X., XU, M. and ZHAO, P. (2023). Modeling multivariate cyber risks: Deep learning dating extreme value theory. *J*. *Appl*. *Stat*. **50** 610–630. MR4545049 https://doi.org/10.1080/02664763.2021.1936468

ZHAO, D., TRAORE, I., SAYED, B., LU, W., SAAD, S., GHORBANI, A. and GARANT, D. (2013). Botnet detection based on traffic behavior analysis and flow intervals. *Comput*. *Secur.* **39** 2–16.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3570–3595
https://doi.org/10.1214/24-AOAS1953

DECONVOLUTION ANALYSIS OF SPATIAL TRANSCRIPTOMICS BY MULTIPLICATIVE-ADDITIVE POISSON-GAMMA MODELS

BY YUTONG LUO[1,a], JOAN E. BAILEY-WILSON[2,b], CHRISTOPHER ALBANESE[3,c] AND RUZONG FAN[4,d]

[1]*Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center,* [a]*yl934@georgetown.edu*

[2]*Office of the Scientific Director, National Human Genome Research Institute, National Institutes of Health,* [b]*jebw@mail.nih.gov*

[3]*Department of Oncology and Radiology, Georgetown University Medical Center,* [c]*albanese@georgetown.edu*

[4]*Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center and Center for Genomics and Data Science Research, National Human Genome Research Institute, National Institutes of Health,* [d]*rf740@georgetown.edu*

Understanding cell type composition and gene expression of spatial transcriptomic data is crucial for comprehending phenotypic variability and detecting key factors that influence disease susceptibility of complex traits. Detecting cell type specific expression patterns from spatial transcriptome profiles is important in studying the cellular components and gene expression of individual cell classes and structural architecture. In this paper we develop mixed effect multiplicative-additive Poisson-gamma models to analyze spatial (MAPS) transcriptomics data using cell type-specific gene expressions in single cell RNA-sequenceing (scRNA-seq) data. To build the mixed effect multiplicative-additive Poisson-gamma models, the gene expression counts of spatial transcriptomics data are treated as dependent variables, and the mean and variance parameters of scRNA-seq data are used to construct independent variables to explain the dependent variables on the basis of Poisson-gamma mixture. One novelty of the proposed mixed models is that the variance parameters of scRNA-seq are used to describe the within-cell-type variations or stochasticity. We develop iteratively analytical formulae to estimate the cell type proportions and dispersion parameters. To address the important research problems and help with intensive spatial transcriptomics data analysis, a readily available software, MAPS, is developed to implement the proposed methods. By simulation study and real data analysis, MAPS is found to perform better than or similar to robust cell type decomposition (RCTD), SpatialDWLS (dampened weighted least squares), conditional autoregressive-based deconvolution (CARD), and a Spatially weighted pOissoN-gAmma Regression model (SONAR). Computationally, MAPS is significantly faster than RCTD and SpatialDWLS. MAPS provides a novel way for mapping spatial tissue architecture.

REFERENCES

10X GENOMICS (2020). visium spatial gene expression. Available at https://www.10xgenomics.com/solutions/spatial-gene-expression.

ANDERSSON, A., BERGENSTRAHLE, J., ASP, M., BERGENSTRAHLE, L., JUREK, A., NAVARRO, J. F. and LUNDEBERG, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3** 1–8.

ANDERSSON, A., LARSSON, L., STENBECK, L., SALMÉN, F., EHINGER, A., WU, S. Z., AL-ERYANI, G., RODEN, D., SWARBRICK, A. et al. (2021). Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat. Commun.* **12** 6012. https://doi.org/10.1038/s41467-021-26271-2

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25. MR0972770

Key words and phrases. Spatial transcriptomicsls, scRNA-seq, deconvolution analysis, cell type expression patterns, mixed effect multiplicative-additive Poisson-gamma models.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5** 403–434. MR1092983

ASP, M., BERGENSTRÅHLE, J. and LUNDEBERG, J. (2020). Spatially resolved transcriptomes—next generation tools for tissue exploration. *BioEssays* **42** 1900221.

ASP, M., GIACOMELLO, S., LARSSON, L., WU, C., FURTH, D., QIAN, X., WARDELL, E., CUSTODIO, J., REIMEGARD, J. et al. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell* **179** 1647–1660.

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

BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation. Oxford Studies in Probability* **2**. The Clarendon Press, Oxford University Press, New York. MR1163825

BERGLUND, E., MAASKOLA, J., SCHULTZ, N., FRIEDRICH, S., MARKLUND, M., BERGENSTRÅHLE, J., TARISH, F., TANOGLIDI, A., VICKOVIC, S. et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9** 2419. https://doi.org/10.1038/s41467-018-04724-5

BJORHOLM, S., SVENNING, J.-C., SKOV, F. and BALSLEV, H. (2008). To what extent does Tobler's 1st law of geography apply to macroecology? A case study using American palms (Arecaceae). *BMC Ecol.* **8** 11. https://doi.org/10.1186/1472-6785-8-11

BRADLOW, E. T., HARDIE, B. G. S. and FADER, P. S. (2002). Bayesian inference for the negative binomial distribution via polynomial expansions. *J. Comput. Graph. Statist.* **11** 189–201. MR1937285 https://doi.org/10.1198/106186002317375677

BROWN, A. M., ARANCILLO, M., LIN, T., CATT, D. R., ZHOU, J., LACKEY, E. P., STAY, T. L., ZUO, Z., WHITE, J. J. et al. (2019). Molecular layer interneurons shape the spike activity of cerebellar Purkinje cells. *Sci. Rep.* **9** 1742. https://doi.org/10.1038/s41598-018-38264-1

CABLE, D. M., MURRAY, E., SHANMUGAM, V., ZHANG, S., ZOU, L. S., DIAO, M., CHEN, H., MACOSKO, E. Z., IRIZARRY, R. A. et al. (2022). Cell type-specific inference of differential expression in spatial transcriptomics. *Nat. Methods* **19** 1076–1087.

CABLE, D. M., MURRAY, E., ZOU, L. S., GOEVA, A., MACOSKO, E. Z., CHEN, F. and IRIZARRY, R. A. (2021). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.* **22** 1–10.

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

CHEN, J., LIU, W., LUO, T., YU, Z., JIANG, M., WEN, J., GUPTA, G. P., GIUSTI, P. and LI, Y. (2022). A comprehensive comparisonon cell-type composition inference for spatial transcriptomics data. *Brief. Bioinform.* **1**. https://doi.org/10.1093/bib/bbac245

CHEN, J., LUO, T., JIANG, M., LIU, J., GUPTA, G. P. and LI, Y. (2023). Cell composition inference and identification of layerspecific spatial transcriptional profiles with POLARIS. *Sci. Adv.* **9** eadd9818.

CHEN, K. H., BOETTIGER, A. N., MOFFITT, J. R., WANG, S. and ZHUANG, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348** aaa6090. https://doi.org/10.1126/science.aaa6090

CHEN, W. T., LU, A., CRAESSAERTS, K., PAVIE, B., FRIGERIO, C. S., CORTHOUT, N., QIAN, X., KOVA, J. L., HNEMUND, M. K. et al. (2020). Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. *Cell* **182** 976–991.

COBOS, F., VANDESOMPELE, J., MESTDAGH, P. and DE PRETER, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34** 1969–1979.

CODELUPPI, S., BORM, L. E., ZEISEL, A., MANNO, G. L., VAN LUNTEREN, J. A., SVENSSON, C. I. and LINNARSSON, S. (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15** 932–935. https://doi.org/10.1038/s41592-018-0175-z

DA SILVA, A. R. and RODRIGUES, T. C. V. (2014). Geographically weighted negative binomial regression—incorporating overdispersion. *Stat. Comput.* **24** 769–783. MR3229696 https://doi.org/10.1007/s11222-013-9401-9

DEAN, C., LAWLESS, J. F. and WILLMOT, G. E. (1989). A mixed Poisson-inverse-Gaussian regression model. *Canad. J. Statist.* **17** 171–181. MR1033100 https://doi.org/10.2307/3314846

DONG, R. and YUAN, G. C. (2021). SpatialDWLS: Accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22** 1–10.

DRIES, R., ZHU, Q., DONG, R., ENG, C. L., LI, H., LIU, K., FU, Y., ZHAO, T., SARKAR, A. et al. (2021). Giotto: A toolbox for integrative analysis and visualization of spatial expression data. *Genome Biol.* **22** 1–31.

EDSGÄRD, D., JOHNSSON, P. and SANDBERG, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15** 339–342. https://doi.org/10.1038/nmeth.4634

ELOSUA-BAYES, M., NIETO, P., MEREU, E., GUT, I. and HEYN, H. (2021). SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49** e50.

ENG, C.-H. L., LAWSON, M., ZHU, Q., DRIES, R., KOULENA, N., TAKEI, Y., YUN, J., CRONIN, C., KARP, C. et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568** 235–239. https://doi.org/10.1038/s41586-019-1049-y

FOTHERINGHAM, A. S., BRUNSDON, C. and CHARLTON, M. (2002). *Geographically Weighted Regression*: *The Analysis of Spatially Varying Relationships* **13**. Wiley, New York.

GREENWOOD, M. and YULE, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. R. Stat. Soc., A* **83** 255–279.

HAGHVERDI, L., LUN, A. T. L., 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. https://doi.org/10.1038/nbt.4091

HUNTER, M. V., MONCADA, R., WEISS, J. M., YANAI, I. and WHITE, R. M. (2021). Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun.* **12** 6278. https://doi.org/10.1038/s41467-021-26614-z

JI, A. L., RUBIN, A. J., THRANE, K., JIANG, S., REYNOLDS, D. L., MEYERS, R. M., GUO, M. G., GEORGE, B. M., MOLLBRINK, A. et al. (2020). Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182** 497–514.

KLESHCHEVNIKOV, V., SHMATKO, A., DANN, E., AIVAZIDIS, A., KING, H. W., LI, T., ELMENTAITE, R., LOMAKIN, A., KEDLIAN, V. et al. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40** 661–671. https://doi.org/10.1038/s41587-021-01139-4

KOZAREVA, V., MARTIN, C., OSORNO, T., RUDOLPH, S., GUO, C., VANDERBURG, C., NADAF, N., REGEV, A., REGEHR, W. G. et al. (2021). A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. *Nature* **598** 214–219.

LANGE, K. (2010). *Numerical Analysis for Statisticians*, 2nd ed. *Statistics and Computing*. Springer, New York. MR2655999 https://doi.org/10.1007/978-1-4419-5945-4

LAWLESS, J. F. (1987). Negative binomial and mixed Poisson regression. *Canad. J. Statist.* **15** 209–225. MR0926553 https://doi.org/10.2307/3314912

LI, B., SEVERSON, E., PIGNON, J., ZHAO, H., LI, T., NOVAK, J., JIANG, P. et al. (2016). Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.* **17** 174.

LI, B., ZHANG, W., GUO, C., XU, H., LI, L., FANG, M., HU, Y. et al. (2022). Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nat. Methods* **19** 174662–670.

LIU, Y., LI, N., QI, J., XU, G., ZHAO, J., WANG, N., HUANG, X. et al. (2023a). A hybrid machine learning and regression method for cell type deconvolution of spatial barcoding-based transcriptomic data. BioRxiv preprint. https://doi.org/10.1101/2023.08.24.554722

LIU, Z., WU, D., ZHAI, W. and MA, L. (2023b). SONAR enables cell type deconvolution with spatially weighted Poisson-Gamma model for spatial transcriptomics. *Nat. Commun.* **14** 4727.

LOPEZ, R., LI, B., KEREN-SHAUL, H., BOYEAU, P., KEDMI, M., PILZER, D., JELINSKI, A. et al. (2022). DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat. Biotechnol.* **40** 1360–1369.

LUO, Y., BAILEY-WILSON, J. E., ALBANESE, C. and FAN, R. (2024). Supplement to "Deconvolution Analysis of Spatial Transcriptomics by Multiplicative-Additive Poisson-gamma Models." https://doi.org/10.1214/24-AOAS1953SUPPA, https://doi.org/10.1214/24-AOAS1953SUPPB

LUO, Y. and FAN, R. (2022). Deconvolution analysis of cell-type expression from bulk tissues by integrating with single-cell expression reference. *Genet. Epidemiol.* **46** 615–628.

MA, Y. and ZHOU, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.* **40** 1349–1359.

MARX, V. (2021). Method of the year 2020: Spatially resolved transcriptomics. *Nat. Methods* **18** 9–14.

MAYNARD, K. R., COLLADO-TORRES, L., WEBER, L. M., UYTINGCO, C., BARRY, B. K., WILLIAMS, S. R., CATALLINI, J. L., TRAN, M. N. II, BESICH, Z. et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24** 425–436.

MOFFITT, J. R., BAMBAH-MUKKU, D., EICHHORN, S. W., VAUGHN, E., SHEKHAR, K., PEREZ, J. D., RUBINSTEIN, N. D., HAO, J., REGEV, A. et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**. https://doi.org/10.1126/science.aau5324

MONCADA, R., BARKLEY, D., WAGNER, F., CHIODIN, M., DEVLIN, J. C., BARON, M., HAJDU, C. H., SIMEONE, D. M. and YANAI, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38** 333–342. https://doi.org/10.1038/s41587-019-0392-8

MOSES, L. and PACHTER, L. (2022). Museum of spatial transcriptomics. *Nat. Methods* **19** 534–546.

NAKAYA, T., FOTHERINGHAM, A. S., BRUNSDON, C. and CHARLTON, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Stat. Med.* **24** 2695–2717. MR2196209 https://doi.org/10.1002/sim.2129

NEWMAN, A. M. and ALIZADEH, A. A. (2016). High-throughput genomic profiling of tumor-infiltrating leukocytes. *Curr. Opin. Immunol.* **41** 77–84. https://doi.org/10.1016/j.coi.2016.06.006

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

NEWMAN, A. M., STEEN, C. B., LIU, C. L., GENTLES, A. J., CHAUDHURI, A. A., SCHERER, F., KHODADOUST, M. S., ESFAHANI, M. S., LUCA, B. A. et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37** 773–782. https://doi.org/10.1038/s41587-019-0114-2

PALLA, G., FISCHER, D. S., REGEV, A. and THEIS, F. J. (2022). Spatial components of molecular tissue biology. *Nat. Biotechnol.* **40** 308–318. https://doi.org/10.1038/s41587-021-01182-1

ROBINSON, M. D. and SMYTH, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9** 321–332.

RODRIQUES, S. G., STICKELS, R. R., GOEVA, A., MARTIN, C. A., MURRAY, E., VANDERBURG, C. R., WELCH, J., CHEN, L. M., CHEN, F. et al. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363** 1463–1467.

SAKAMOTO, Y., ISHIGURO, M. and KITAGAWA, G. (1986). *Akaike Information Criterion Statistics*. *Mathematics and Its Applications* (*Japanese Series*) **1**. D. Reidel Publishing, Dordrecht. MR0876486

SARKAR, A. and STEPHENS, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53** 770–777. https://doi.org/10.1038/s41588-021-00873-4

SAUNDERS, A., MACOSKO, E. Z., WYSOKER, A., GOLDMAN, M., KRIENEN, F. M., DE RIVERA, H., BIEN, E., BAUM, M., BORTOLIN, L. et al. (2018). Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174** 1015–1030.

SONG, Q. and SU, J. (2021). DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Brief. Bioinform.* **22** 1–13.

STÅHL, P. L., SALMÉN, F., VICKOVIC, S., LUNDMARK, A., NAVARRO, J. F., MAGNUSSON, J., GIACOMELLO, S., ASP, M., WESTHOLM, J. O. et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353** 78–82. https://doi.org/10.1126/science.aaf2403

STEEN, C. B., LIU, L. C., ALIZADEH, A. A. and NEWMAN, A. M. (2020). Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. *Stem. Cell. Transcriptional Networks* **2117** 135–157.

STICKELS, R. R., MURRAY, E., KUMAR, P., LI, J., MARSHALL, J. L., BELLA, D. J. D., ARLOTTA, P., MACOSKO, E. Z. and CHEN, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39** 313–319.

STUART, T. and SATIJA, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* **20** 257–272. https://doi.org/10.1038/s41576-019-0093-7

TOBLER, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46** 234–240.

TSOUCAS, D., DONG, R., CHEN, H., ZHU, Q., GUO, G. and YUAN, G. C. (2019). Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10** 2975.

VICKOVIC, S., ERASLAN, G., SALMEN, F., KLUGHAMMER, J., STENBECK, L., SCHAPIRO, D., AIJO, T., BONNEAU, R., BERGENSTRAHLE, L. et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16** 987–990.

WANG, X., PARK, J., SUSZTAK, K., ZHANG, N. R. and LI, M. Y. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10** 380.

ZHOU, M., LI, L., DUNSON, D. and CARIN, L. (2012). Lognormal and gamma mixed negative binomial regression. In *Proc. Int. Conf. Mach. Learn.* 1343–1350.

The Annals of Applied Statistics
2024, Vol. 18, No. 4, 3596–3613
https://doi.org/10.1214/24-AOAS1954

DEEPMAP: DEEP LEARNING-BASED SINGLE-CELL DATA INTEGRATION USING ITERATIVE CELL MATCHING AND STRUCTURE PRESERVATION CONSTRAINTS

BY SHUNTUO XU[1,a], ZHOU YU[1,b] AND JINGSI MING[2,c]

[1]*KLATASDS-MOE, School of Statistics, East China Normal University,* [a]*oaksword@163.com,* [b]*zyu@stat.ecnu.edu.cn*
[2]*Academy of Statistics and Interdisciplinary Sciences, East China Normal University,* [c]*jsming@fem.ecnu.edu.cn*

Effective integration of single-cell data can facilitate the discovery of cell-type specific gene expression patterns and cellular interactions, ultimately leading to a better understanding of various biological processes and diseases. However, datasets from different platforms, species, and modalities exhibit various levels of heterogeneities, posing significant challenges in data alignment using a unified approach. Here we propose DeepMap, a flexible and efficient method for single-cell data integration, by taking advantage of the deep learning framework. Our method utilizes iterative cell matching based on mutual nearest neighbors, leverages an autoencoder framework to learn harmonized representations of cells from various datasets, and incorporates a covariance penalty term into the framework for structure preservation. In addition to harmonization of data from different datasets, we specifically take account of the preservation of important biological variations within dataset, which is crucial to reliable downstream analysis. Comprehensive real data analysis demonstrates the flexibility of DeepMap for diverse datasets from different platforms, species, and modalities, and highlights its marked ability in preserving structures over existing integration methods with enhanced computational efficiency and optimized memory usage. The robust DeepMap-integrated data offers promising prospects for advancing our understanding of cell biology, hence making it a highly attractive option for integrative single-cell data analysis.

REFERENCES

BAKKEN, T. E., HODGE, R. D., MILLER, J. A., YAO, Z., NGUYEN, T. N., AEVERMANN, B., BARKAN, E., BERTAGNOLLI, D., CASPER, T. et al. (2018). Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS ONE* **13** e0209648.

BAKKEN, T. E., JORSTAD, N. L., HU, Q., LAKE, B. B., TIAN, W., KALMBACH, B. E., CROW, M., HODGE, R. D., KRIENEN, F. M. et al. (2021). Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598** 111–119.

BARON, M., VERES, A., WOLOCK, S. L., FAUST, A. L., GAUJOUX, R., VETERE, A., RYU, J. H., WAGNER, B. K., SHEN-ORR, S. S. et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst*. **3** 346–360.

CAO, J., SPIELMANN, M., QIU, X., HUANG, X., IBRAHIM, D. M., HILL, A. J., ZHANG, F., MUNDLOS, S., CHRISTIANSEN, L. et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566** 496–502.

CAO, Z.-J. and GAO, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol*. **40** 1458–1466.

CHEN, C., XING, D., TAN, L., LI, H., ZHOU, G., HUANG, L. and XIE, X. S. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356** 189–194.

CHEN, Y., ZHENG, Y., GAO, Y., LIN, Z., YANG, S., WANG, T., WANG, Q., XIE, N., HUA, R. et al. (2018). Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res*. **28** 879–896.

Key words and phrases. Single-cell data integration, iterative cell matching, structure preservation constraints, deep learning.

Crinier, A., Milpied, P., Escaliere, B., Piperoglou, C., Galluso, J., Balsamo, A., Spinelli, L., Cervera-Marzal, I., Ebbo, M. et al. (2018). High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice. *Immunity* **49** 971–986.

Ginhoux, F., Yalin, A., Dutertre, C. A. and Amit, I. (2022). Single-cell immunology: Past, present, and future. *Immunity* **55** 393–404. https://doi.org/10.1016/j.immuni.2022.02.006

Green, C. D., Ma, Q., Manske, G. L., Shami, A. N., Zheng, X., Marini, S., Moritz, L., Sultan, C., Gurczynski, S. J. et al. (2018). A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq. *Dev. Cell* **46** 651–667.

Grün, D., Muraro, M. J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., Van Es, J., Jansen, E. et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19** 266–277.

Haghverdi, L., Lun, A. T. L., 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. https://doi.org/10.1038/nbt.4091

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H. et al. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell* **172** 1091–1107.

Hermann, B. P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K. N., Chen, I.-C., Gildersleeve, H., Lehle, J. D., Mayo, M. et al. (2018). The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Rep.* **25** 1650–1667.

Hie, B., Bryson, B. and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37** 685–691. https://doi.org/10.1038/s41587-019-0113-3

Howick, V. M., Russell, A. J., Andrews, T., Heaton, H., Reid, A. J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L. H. et al. (2019). The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science* **365**. eaaw2619.

Hu, J., Schroeder, A., Coleman, K., Chen, C., Auerbach, B. J. and Li, M. (2021). Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Comput. Struct. Biotechnol. J.* **19** 3829–3841.

Iram, T., Consortium, T. M. et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562** 367–372.

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.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P. and Stitzel, M. L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27** 208–222. https://doi.org/10.1101/gr.212720.116

Li, H., Li, H., Zhou, J. and Gao, X. (2022). SD2: Spatially resolved transcriptomics deconvolution through integration of dropout and spatial information. *Bioinformatics* **38** 4878–4884.

Lin, Y., Wu, T.-Y., Wan, S., Yang, J. Y., Wong, W. H. and Wang, Y. R. (2022). scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.* **40** 703–710.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15** 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Lu, Y., Shiau, F., Yi, W., Lu, S., Wu, Q., Pearson, J. D., Kallman, A., Zhong, S., Hoang, T. et al. (2020). Single-cell analysis of human retina identifies evolutionarily conserved and species-specific mechanisms controlling development. *Dev. Cell* **53** 473–491.

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, R., Sun, E. D., Donoho, D. and Zou, J. (2024). Principled and interpretable alignability testing and integration of single-cell data. *Proc. Natl. Acad. Sci. USA* **121**. e2313719121.

Malkov, Y. A. and Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Marder, E. and Goaillard, J.-M. (2006). Variability, compensation and homeostasis in neuron and network function. *Nat. Rev. Neurosci.* **7** 563–574. https://doi.org/10.1038/nrn1949

Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B. Z. et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39** 1246–1258. https://doi.org/10.1038/s41587-021-00927-2

MING, J., LIN, Z., ZHAO, J., WAN, X., THE TABULA MICROCEBUS CONSORTIUM, YANG, C. and WU, A. R. (2022). FIRM: Flexible integration of single-cell RNA-sequencing data for large-scale multi-tissue cell atlas datasets. *Brief. Bioinform.* **23** bbac167.

MING, J., ZHAO, J. and YANG, C. (2023). scPI: A scalable framework for probabilistic inference in single-cell RNA-sequencing data analysis. *Stat. Biosci.* **15** 633–656.

MOFFITT, J. R., BAMBAH-MUKKU, D., EICHHORN, S. W., VAUGHN, E., SHEKHAR, K., PEREZ, J. D., RUBINSTEIN, N. D., HAO, J., REGEV, A. et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**. eaau5324.

MURARO, M. J., DHARMADHIKARI, G., GRÜN, D., GROEN, N., DIELEN, T., JANSEN, E., VAN GURP, L., ENGELSE, M. A., CARLOTTI, F. et al. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3** 385–394.

MUSTACHIO, L. M. and ROSZIK, J. (2022). Single-cell sequencing: Current applications in precision oncogenomics and cancer therapeutics. *Cancers* (*Basel*) **14** 657. https://doi.org/10.3390/cancers14030657

NADELMANN, E. R., GORHAM, J. M., REICHART, D., DELAUGHTER, D. M., WAKIMOTO, H., LINDBERG, E. L., LITVIŇUKOVA, M., MAATZ, H., CURRAN, J. J. et al. (2021). Isolation of nuclei from mammalian cells and tissues for single-nucleus molecular profiling. *Curr. Protoc.* **1** e132.

NGUYEN, Q. H., PERVOLARAKIS, N., BLAKE, K., MA, D., DAVIS, R. T., JAMES, N., PHUNG, A. T., WILLEY, E., KUMAR, R. et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**. 2028.

NOMURA, S. (2021). Single-cell genomics to understand disease pathogenesis. *J. Hum. Genet.* **66** 75–84.

QIU, X., MAO, Q., TANG, Y., WANG, L., CHAWLA, R., PLINER, H. A. and TRAPNELL, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14** 979–982. https://doi.org/10.1038/nmeth.4402

REN, L., WANG, J., LI, Z., LI, Q. and YU, G. (2023). scMCs: A framework for single-cell multi-omics data integration and multiple clusterings. *Bioinformatics* **39** btad133.

ROSENBERG, A. B., ROCO, C. M., MUSCAT, R. A., KUCHINA, A., SAMPLE, P., YAO, Z., GRAYBUCK, L. T., PEELER, D. J., MUKHERJEE, S. et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360** 176–182.

SAUNDERS, A., MACOSKO, E. Z., WYSOKER, A., GOLDMAN, M., KRIENEN, F. M., DE RIVERA, H., BIEN, E., BAUM, M., BORTOLIN, L. et al. (2018). Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174** 1015–1030.

SEGERSTOLPE, Å., PALASANTZA, A., ELIASSON, P., ANDERSSON, E.-M., ANDRÉASSON, A.-C., SUN, X., PICELLI, S., SABIRSH, A., CLAUSEN, M. et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24** 593–607.

SHAMI, A. N., ZHENG, X., MUNYOKI, S. K., MA, Q., MANSKE, G. L., GREEN, C. D., SUKHWANI, M., ORWIG, K. E., LI, J. Z. et al. (2020). Single-cell RNA sequencing of human, macaque, and mouse testes uncovers conserved and divergent features of mammalian spermatogenesis. *Dev. Cell* **54** 529–547.

STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK, W. M., HAO, Y., STOECKIUS, M., SMIBERT, P. et al. (2019). Comprehensive integration of single-cell data. *Cell* **177** 1888–1902.

TIAN, L., CHEN, F. and MACOSKO, E. Z. (2023). The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.* **41** 773–782. https://doi.org/10.1038/s41587-022-01448-2

VITAK, S. A., TORKENCZY, K. A., ROSENKRANTZ, J. L., FIELDS, A. J., CHRISTIANSEN, L., WONG, M. H., CARBONE, L., STEEMERS, F. J. and ADEY, A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14** 302–308. https://doi.org/10.1038/nmeth.4154

WANG, L., YANG, Y., MA, H., XIE, Y., XU, J., NEAR, D., WANG, H., GARBUTT, T., LI, Y. et al. (2022). Single-cell dual-omics reveals the transcriptomic and epigenomic diversity of cardiac non-myocytes. *Cardiovasc. Res.* **118** 1548–1563.

WELCH, J. D., KOZAREVA, V., FERREIRA, A., VANDERBURG, C., MARTIN, C. and MACOSKO, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177** 1873–1887.

WOLF, F. A., ANGERER, P. and THEIS, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19** 1–5.

WU, F., FAN, J., HE, Y., XIONG, A., YU, J., LI, Y., ZHANG, Y., ZHAO, W., ZHOU, F. et al. (2021). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* **12** 2540.

WU, H., KIRITA, Y., DONNELLY, E. L. and HUMPHREYS, B. D. (2019). Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: Rare cell types and novel cell states revealed in fibrosis. *J. Amer. Soc. Nephrol.* **30** 23–32. https://doi.org/10.1681/ASN.2018090912

XIA, C., FAN, J., EMANUEL, G., HAO, J. and ZHUANG, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. USA* **116** 19490–19499.

XU, S., YU, Z. and MING, J. (2024). Supplement to "DeepMap: Deep learning-based single-cell data integration using iterative cell matching and structure preservation constraints." https://doi.org/10.1214/24-AOAS1954SUPPA, https://doi.org/10.1214/24-AOAS1954SUPPB, https://doi.org/10.1214/24-AOAS1954SUPPC

YAZAR, S., ALQUICIRA-HERNANDEZ, J., WING, K., SENABOUTH, A., GORDON, M. G., ANDERSEN, S., LU, Q., ROWSON, A., TAYLOR, T. R. et al. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**. eabf3041.

ZEISEL, A., HOCHGERNER, H., LÖNNERBERG, P., JOHNSSON, A., MEMIC, F., VAN DER ZWAN, J., HÄRING, M., BRAUN, E., BORM, L. E. et al. (2018). Molecular architecture of the mouse nervous system. *Cell* **174** 999–1014.

ZENG, Z., LI, Y., LI, Y. and LUO, Y. (2022). Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol*. **23** 1–23.

ZETHOVEN, M., MARTELOTTO, L., PATTISON, A., BOWEN, B., BALACHANDER, S., FLYNN, A., ROSSELLO, F. J., HOGG, A., MILLER, J. A. et al. (2022). Single-nuclei and bulk-tissue gene-expression analysis of pheochromocytoma and paraganglioma links disease subtypes with tumor microenvironment. *Nat. Commun*. **13** 6262.

ZHANG, L., YU, X., ZHENG, L., ZHANG, Y., LI, Y., FANG, Q., GAO, R., KANG, B., ZHANG, Q. et al. (2018). Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564** 268–272.

ZHAO, J., WANG, G., MING, J., LIN, Z., WANG, Y., WU, A. R. and YANG, C. (2022). Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. *Nat. Comput. Sci*. **2** 317–330.

ZOU, B., ZHANG, T., ZHOU, R., JIANG, X., YANG, H., JIN, X. and BAI, Y. (2021). deepMNN: Deep learning-based single-cell RNA sequencing data batch correction using mutual nearest neighbors. *Front. Genet*. **12** 708981.